A Context-Sensitive Generalization of ICA

Barak A. Pearlmutter† Lucas C. Parra‡
†Dept. of Cog. Sci., UCSD, La Jolla, California, USA, barak.pearlmutter@alumni.cs.cmu.edu
‡Siemens Corporate Research, Princeton, New Jersey, USA, lucas@scr.siemens.com

Abstract— **Source separation arises in a surprising number of signal processing applications, from speech recognition to EEG analysis. In the square linear blind source separation problem without time delays, one must find an unmixing matrix which can detangle the result of mixing n unknown independent sources through an unknown $n \times n$ mixing matrix. The recently introduced ICA blind source separation algorithm (Baram and Roth 1994; Bell and Sejnowski 1995) is a powerful and surprisingly simple technique for solving this problem. ICA is all the more remarkable for performing so well despite making absolutely no use of the temporal structure of its input! This paper presents a new algorithm, contextual ICA, which derives from a maximum likelihood density estimation formulation of the problem. cICA can incorporate arbitrarily complex adaptive history-sensitive source models, and thereby make use of the temporal structure of its input. This allows it to separate in a number of situations where standard ICA cannot, including sources with low kurtosis, colored gaussian sources, and sources which have gaussian histograms. Since ICA is a special case of cICA, the MLE derivation provides as a corollary a rigorous derivation of classic ICA.**

1 The ICA algorithm

In the blind source separation problem, one is given the output of a number of microphones, each of which records a mixture of a number of sources. The task is to recover the sources. In the blind linear square case, there are the same number of microphones as sources, and the mixing is linear. In the absence of time delays or echos, the mixing is characterized by an $n \times n$ matrix $\mathbf{A}$, so if $\mathbf{s}(t)$ is a vector of the sources at time t then $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ is a vector of the signals received by the microphones at time t. Naturally we will assume that $\mathbf{A}$ is full rank.

In the absence of noise, which is the case we consider, the solution to this problem is to find a full rank $n \times n$ matrix $\mathbf{W}$ which has the property that $\mathbf{WA}$ has exactly one nonzero element in each row and each column. We denote the result of the unmixing process as $\mathbf{y}(t)$, and note that $\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) = \mathbf{WA}\mathbf{s}(t)$. If we have found an appropriate $\mathbf{W}$ then the product $\mathbf{WA}$ will be equal to the product of a diagonal matrix with a permutation matrix, and the elements of $\mathbf{y}(t)$ will be the same as the elements of $\mathbf{s}(t)$, but shuffled and scaled.

With no prior information about $\mathbf{A}$ or the source signals $s_i(t)$, the problem might sound impossible. However, for non-gaussian distributions, it is not. An algorithm called *independent components analysis* was introduced by Comon (1994). This version of the algorithm approximates some distributions by their first few moments, which is both approximate and computationally burdensome. Single coordinate higher order cumulants are used in a somewhat simpler algorithm by Obradovic and Deco (1995). A surprisingly simple, but inexpensive and exact, variant of the Comon (1994) algorithm was recently introduced (Baram and Roth 1994; Bell and Sejnowski 1995). In a now standard abuse of notation, this new algorithm will be refered to as ICA. This simpler ICA algorithm takes each component of the vector $\mathbf{y}(t)$ and passes it though a saturating monotonic nonlinearity, giving a vector $\mathbf{z}(t)$. Gradient descent is used to modify the components of the matrix $\mathbf{W}$ and the bias terms of the nonlinearities in order to increase the entropy of the distribution of $\mathbf{z}(t)$ induced by the input distribution. ICA was motivated by considerations of biological optimality, which flow from experiments showing that, when presented with natural stimuli, many neurons appear to make good use of their available axonal channel capacity (Bialek *et al.* 1991).

The ICA algorithm, in various configurations, has been applied to a surprising number of problems, from separation of digitally mixed speech signals (Bell and Sejnowski 1995), to separating the componenets of electroencephalographic data (Makeig *et al.* 1996), to blind deconvolution (Bell and Sejnowski 1995), to finding the higher-order structure of a natural sound (Bell and Sejnowski 1996b), and even to financial forecasting (Baram and Roth 1995) and image processing (Bell and Sejnowski 1996a). There have been attempts to generalize the algorithm, the most notable being extensions to tolerate time delays and echos introduced by Torkkola (1996a, 1996b).

The usual intuition for why ICA tends to separate sources runs roughly as follows: if the output entropy is maximized, then the components of the output vector must be statistically independent. If so, then the signals must also be statistically independent prior to the nonlinearity. That being the case, the sources must be separated.

However, there are problematic cases which ICA cannot separate. For instance, a mixture of two uniform distributions, or more generally two low-kurtosis distributions, is not properly separated. (Although separation in this case might be achieved by using a special nonlinearity chosen for the problem.) Since a two-dimensional gaussian distribution is rotationally symmetric, a mixture of white gaussian sources is inherently impossible to separate. Any

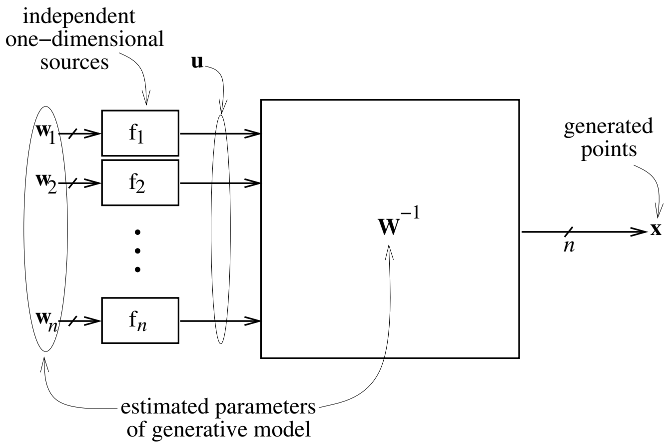

Figure 1: The ICA algorithm fits this parameterized generative model to data.

algorithm that makes no use of the temporal structure of its inputs can by definition make use of only the cumulative histograms of its inputs. If these histograms are gaussian, then such an algorithm will be in principle unable to separate. Since ICA makes no use of the temporal structure of its inputs, it is in principle unable to separate sources whose histograms are gaussian. This includes, for example, colored gaussian sources, speech or music which happen to have gaussian histograms, etc. It is sometimes speculated that any mixture of sources with high-kurtosis histograms is separable by ICA—but there is as yet no proof of this.

We shall now proceed to derive an ICA-like algorithm that can make use of temporal context. We do this by reformulating the blind source separation problem in a maximum likelihood framework.

2 Source separation and maximum likelihood density estimation

Consider the abstract problem of density estimation from samples. One desires to estimate some true distribution $p(\mathbf{x})$ over a space $\mathcal{R}^n$ from which samples $\mathbf{x}_1, \mathbf{x}_2, \ldots$ have been drawn. The maximum likelihood approach (Mendel and Burrus 1990) is to use a density estimator of some parametric form, say $\hat{p}(\mathbf{x};\mathbf{w})$. Given a setting of the parameter vector $\mathbf{w}$, this will constitute the estimated probability density. In order to set $\mathbf{w}$ appropriately, we find a value for it that minimizes a measure of the difference between $p(\mathbf{x})$ and $\hat{p}(\mathbf{x};\mathbf{w})$. An appropriate difference measure is the asymmetric divergence

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

This is the entropy of the (fixed) input distribution p minus the likelihood of p given $\hat{p}$, and the $\mathbf{w}$ which minimizes this maximizes the likelihood; hence the term. (In a full Bayesian treatment, a prior distribution over $\hat{p}$ would have to be specified. This term would manifest itself here as an extra term giving the description length of the model $\hat{p}$.)

Although G itself is not available to us, an unbiased estimate of it can be obtained by taking a sample $\mathbf{x}$ from p,

$$\widehat{G} = H[p] - \log \hat{p}(\mathbf{x};\mathbf{w}) \tag{2}$$

In order to apply a stochastic gradient optimization method, we wish to find an unbiased estimate of $dG/d\mathbf{w}$ (Robbins and Monro 1951). Due to the linearity of differentiation, $d\widehat{G}/d\mathbf{w} = -(d/d\mathbf{w}) \log \hat{p}(\mathbf{x};\mathbf{w})$ is such an estimate.

For blind source separation, we consider the parametric form for $\hat{p}(\mathbf{x};\mathbf{w})$ shown in figure 1. Let $\mathbf{u}$ be an n-dimensional vector whose components u_j are drawn from n independent parameterized one-dimensional density functions $f_j(u_j;\mathbf{w}_j)$. Now let $\mathbf{W}$ be an $n \times n$ matrix, and let $\mathbf{x} = \mathbf{W}^{-1}\mathbf{u}$. The consequent density on $\mathbf{x}$ is denoted $\hat{p}(\mathbf{x};\mathbf{w})$, where the parameter vector $\mathbf{w}$ is a concatenation of the elements of $\mathbf{W}$ with the parameters $\mathbf{w}_1, \ldots, \mathbf{w}_n$ of the densities $f_1, \ldots, f_n$. The components of $\mathbf{u}$ represent the n independent sources which we would like to recover from the observed linear mix $\mathbf{x}$, and $\mathbf{W}$ represents the appropriate unmixing matrix.

To calculate $d\widehat{G}/d\mathbf{w}$ we expand $\log \hat{p}(\mathbf{x};\mathbf{w}) = \log|\mathbf{W}| + \sum_j \log f_j(u_j;\mathbf{w}_j)$ where $\mathbf{u} = \mathbf{W}\mathbf{x}$. We then obtain

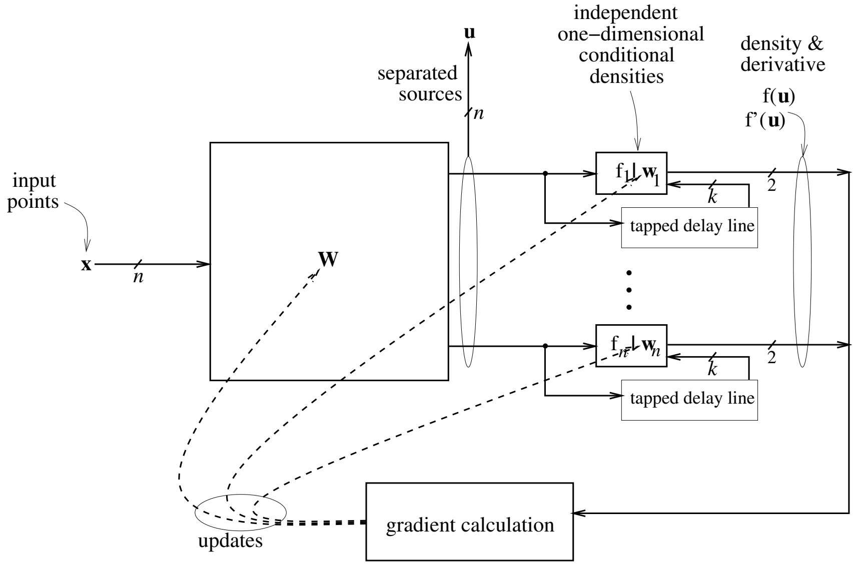

Figure 2: The contextual ICA (cICA) algorithm uses conditional densities which are not memoryless.

formulas for the two different sorts of parameters involved,

$$\frac{d\widehat{G}}{d\mathbf{W}} = -\mathbf{W}^{-T} - \left(\frac{f_j'(u_j;\mathbf{w}_j)}{f_j(u_j;\mathbf{w}_j)}\right)_j \mathbf{x}^T \tag{3}$$

$$\frac{d\widehat{G}}{d\mathbf{w}_j} = -\frac{df_j(u_j;\mathbf{w}_j)/d\mathbf{w}_j}{f_j(u_j;\mathbf{w}_j)} \tag{4}$$

where $(\mathit{expr}(j))_j$ denotes the column vector whose elements are $\mathit{expr}(1),\ldots,\mathit{expr}(n)$.

This is precisely the ICA algorithm, where our $f_j(u_j;\mathbf{w}_j)$ is the derivative of the Bell and Sejnowski (1995) saturating monotonic nonlinearity $g(u_j)$, and our parameter vector $\mathbf{w}_j$ holds the j^{th} component of their $\mathbf{w}_0$ vector of bias terms, $f_j(u_j;\mathbf{w}_j) = g'(u_j + (\mathbf{w}_0)_j)$. In our formulation no squashing nonlinearity is ever calculated, except perhaps as a common subexpression in the computation of the derivatives of the densities. However, the output of the squashing nonlinearity is never actually used for anything in classic ICA.

3 Generalizing ICA

Under this MLE formulation of source separation, there is no restriction on the form of the distributions f_j. The density function $f_j(u_j)$ can have complex structure, and can be conditioned on other information—such as its recent history (as shown in figure 2), or even information from other modalities. All that is required is that the components of $\mathbf{u}$ be *conditionally* independent. In general, f_j can be of the form

$$f_j(u_j(t)|\mathbf{u}(t-1),\mathbf{u}(t-2),\ldots,\text{other information},\ldots;\mathbf{w}_j)$$

We call this algorithm *contextual ICA* or cICA. To give a vivid example, if the sources were different people speaking, then the "other information" might be lip position measured using a visual modality, and $u_j(t)$ would be primarily conditioned on the recent history of that source itself, $u_j(t-1), u_j(t-2),\ldots$, but there might also be some small influence from other speakers. Although f_j can in principle be made arbitrarily complex, there is no practical reason to make it more complex than is necessary to permit proper separation of the sources.

Of course we must still calculate $df_j(u_j;\mathbf{w}_j)/d\mathbf{w}_j$ as per equation 4. In doing so, the history $u_j(t-1), u_j(t-2),\ldots$ of source j is treated as constant with respect to changes in $\mathbf{w}_j$. This is correct, because the unmixing depends only on the matrix $\mathbf{W}$ and not the parameters $\mathbf{w}_j$ of the individual source distributions. On the other hand, changing $\mathbf{W}$ changes the estimated recent history of source u_j, which in turn has an influence on f_j. However we use equation 3 without adding these extra terms. The approximation of dropping these cross terms is ubiquitous in time series analysis, and in this case the successful results of our simulations leads us to believe that it is benign.

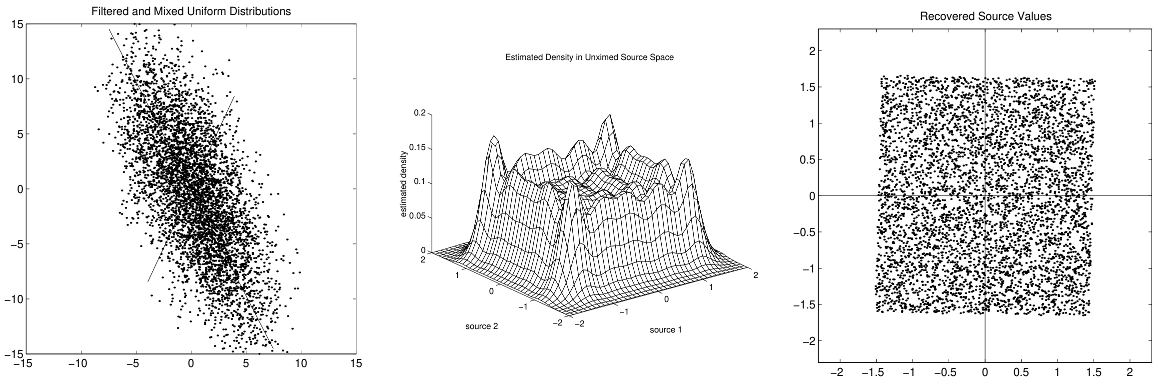

Figure 3: cICA using a history of one time step and a mixture of five logistic densities for each source was applied to 5,000 samples of a mixture of two one-dimensional uniform distributions each filtered by convolution with a decaying exponential of time constant of 99.5. Shown is a scatterplot of the data input to the algorithm, along with the true source axes (left), the estimated residual probability density (center), and a scatterplot of the residuals of the data transformed into the estimated source space coordinates (right). The product of the true mixing matrix and the estimated unmixing matrix deviates from a scaling and permutation matrix by about 3%.

4 Experiments

In our simulations we chose to make f_j a weighted sum of logistic density functions[1] with variable means and scales, and make these means linear functions of the recent history of source j. This allowed us to revert to classic ICA by setting the amount of temporal context to zero and the number of logistic densities in the sum to one. This density estimator, and the corresponding derivatives, are described in detail in appendix A.

Here we experiment with two distributions that conventional ICA is unable to separate. The first is an extremely simple two-dimensional distribution with no temporal context: both x_1 and x_2 are chosen iid from a uniform distribution. Conventional ICA incorrectly rotates the distribution 45 degrees, for reasons explained very well by Bell and Sejnowski (1995) in their discussion of this problematic case. The cICA algorithm successfully separates the sources. To make the problem more challenging, we then filtered each source through low-pass filter. The resulting time series has very gaussian histograms, but as shown in figure 3, cICA again correctly separates the sources.

The second experiment is somewhat more involved. Ten acoustic sources, which include the six used by Bell and Sejnowski (1995), were obtained, courtesty of Dr. Tony Bell. As shown in figure 4, the cumulative density of each source was measured and used to construct a monotonically increasing normalizer which, when applied to each sample from a source, gave the time series a gaussian histogram. These preprocessed time series were mixed using a random matrix. As shown in figures 5 and 6, ICA was unable to separate the resulting babble, but cICA separates properly, even when using only a very small amount of temporal context.

5 Discussion

In deriving cICA we have seen that ICA can regarded as a gradient method for performing maximum likelihood density estimation using a linear historyless factorial model and rigid source densities. The resulting error measure is naturally the same as in the Bell and Sejnowski (1995) derivation, but taking an MLE viewpoint allows a number of generalizations, which allow cICA to to separate a wider variety of sources.

A weakness this method shares with other blind source separation techniques is that it not robust to modulation of the dimensionality. In other words, it is not designed for a non-square mixing matrix. If $\mathbf{x} = \mathbf{A}\mathbf{s}$ and $\mathbf{x}$ is n-dimensional but $\mathbf{s}$ is m-dimensional, then in the case that $n > m$ the algorithm presented here can make no good use of the extra information but to imagine that a few extra Gaussian sources were mixed into the signal. This may perhaps be solved by using a $\mathbf{W}$ matrix of a special form. In the case that $n < m$ no linear unmixing can separate the sources, and it seems that a strong prior will be necessary to distinguish a single complex one-dimensional source from the one-dimensional sum of two simple independent one-dimensional sources, and a nonlinear unmixing process will be necessary to separate them.

[1] If $g(t)$ is the fraction of the susceptible population already infected, then the Verhulst (1844) epidemic equation, $dg/dt = g(t)(1 - g(t))$, expresses a random-contact homogeneous-population model of growth. This results in a logistic cumulative distribution function $g(t) = 1/(1+\exp -t)$. The logistic density function is $h(t) = dg/dt$, the corresponding probability density of contracting the disease at time t.

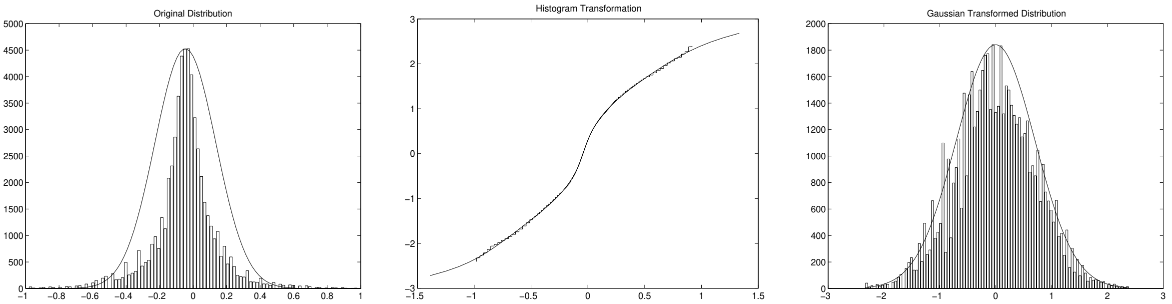

Figure 4: Histogram of samples from one of the acoustic sources used in the mixture below (left), nonlinear transformation applied to the data (center), histogram of transformed data (right).

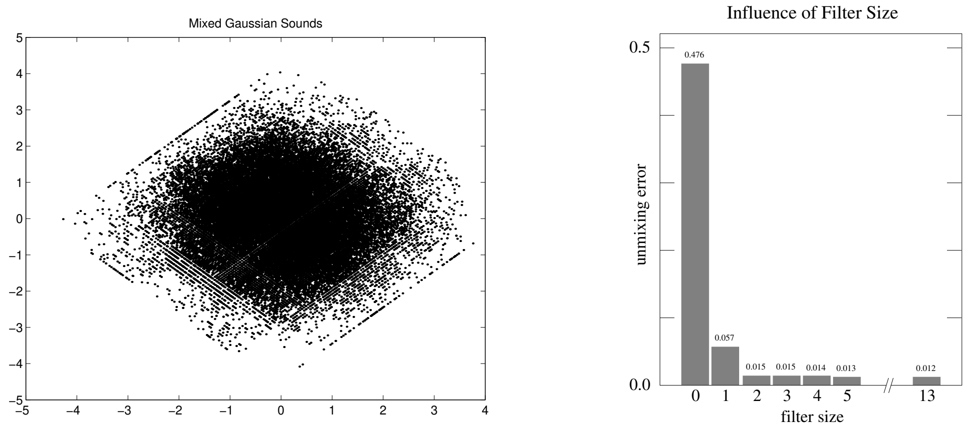

Figure 5: Scatterplot of linear mixture of two gaussianified acoustic sources (left), and unmixing error of cICA (using linear predictive sources with a single logistic) as a function of the length of the history used in the predictive filter (right). The zero history case corresponds to classic ICA, which fail to separate due to the gaussian histograms. (The parallelogram-shaped boundary and the stripes in the scatterplot on the left are artifacts of the signal quantization and the digital mixing.)

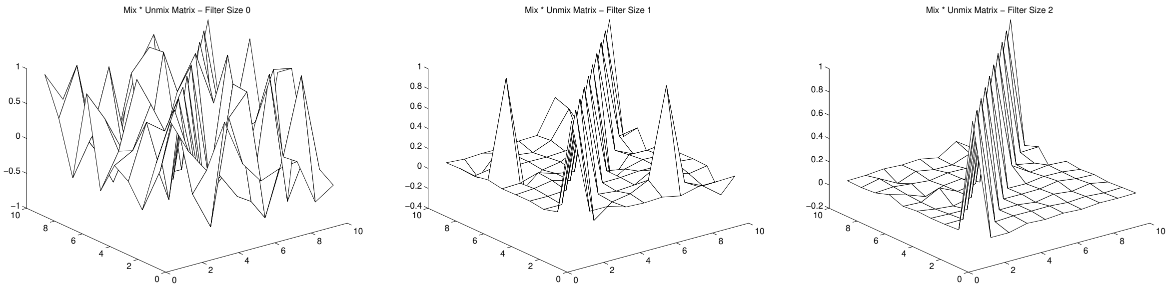

Figure 6: Plot of the elements of the product of the true mixing matrix and the estimated unmixing matrix, with each row normalized to make the largest element equal to one, and the rows permuted to place large elements along the diagonal. If the unmixing is perfect, the result will be a ridge along the diagonal with all off-diagonal elements equal to zero. The ten sources mixed are acoustic sources (courtesy of Tony Bell) which have had a monotonic nonlinearity applied to them to make their histograms exhibit gaussian statistics (see figure 4.) These are mixed using a random mixing matrix, and cICA with linear predictive sources and a single logistic density is used to estimate the unmixing matrix. The length of the history used is varied from zero, which corresponds to conventional ICA (left), to one (center), to two (right).

Finally, we would like to compare ICA with PCA. The principal components algorithm (Hotelling 1933) fits a linear mixture of one-dimensional Gaussian sources of minimal variance to samples from a high-dimensional distribution. ICA performs a similar action, but instead uses a linear mixture of potentially non-Gaussian distributions. As such, ICA might be viewed as a linear but non-Gaussian generalization of PCA—except that without PCA's minimum variance constraint, if Gaussian distributions are used for the f_j distributions of ICA, the unmixing matrix W has a great deal of freedom. It need not be orthogonal, and the coordinate system it embodies need have no special status. A challenge that remains with us is to find a sensible nonlinear analogue of PCA. One algorithm was proposed for this purpose by Parra, Deco, and Miesbach (1995), who replaced the orthogonal linear mixture of PCA by a symplectic mixing function while retaining PCA's minimal variance Gaussian source model. Unfortunately the symplectic map has a great deal of undesired freedom, so again the coordinate system it produces need have no special status.

6 Future work

Our current work concentrates on combining source separation with deconvolution, to enable the system to both tolerate and cancel the effects of echos and time delays between the sources and the microphones. An inherent ambiguity is introduced, which amounts to a freedom of one filter per source. We hope to resolve this ambiguity in a more symmetric fashion than in Torkkola (1996a), where identity filters are placed along the diagonal of the matrix of deconvolution filters. We are also exploring the incorporation of microphone nonlinearities, and microphone noise of known distribution, into the model.

Acknowledgments

Thanks are due to Dr. Tony Bell for provocative discussions and for generously sharing his data. Portions of this work were performed while BAP was visiting the Sloan Center for Theoretical Neurobiology at the Salk Institute.

References

Amari, S., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation.. In NIPS*95 (1996). In press.

Baram, Y. and Roth, Z. (1994). Density Shaping by Neural Networks with Application to Classification, Estimation and Forecasting. Tech. rep. CIS-94-20, Center for Intelligent Systems, Technion, Israel Institute for Technology, Haifa.

Baram, Y. and Roth, Z. (1995). Forecasting by Density Shaping Using Neural Networks. In *Computational Intelligence for Financial Engineering* New York City. IEEE Press.

Bell, A. J. and Sejnowski, T. J. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Bell, A. J. and Sejnowski, T. J. (1996a). The Independent Components of Natural Scenes. *Vision Research*. Submitted.

Bell, A. J. and Sejnowski, T. J. (1996b). Learning the higher-order structure of a natural sound. *Network*. In press.

Bialek, W., Rieke, F., de Ruyter van Stevenick, R. R., and Warland, D. (1991). Reading a Neural Code. *Science*, *252*, 1854–1857.

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

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441, 498–520.

Makeig, S., Bell, A. J., Jung, T.-P., and Sejnowski, T. J. (1996). Independent component analysis of Electroencephalographic data.. In NIPS*95 (1996). In press.

Mendel, J. M. and Burrus, C. S. (1990). *Maximum-likelihood deconvolution: a journey into model-based signal processing*. Springer-Verlag.

NIPS*95 (1996). *Advances in Neural Information Processing Systems 8*. MIT Press. In press.

Nowlan, S. J. and Hinton, G. E. (1992). Adaptive Soft Weight Tying using Gaussian Mixtures. In *Advances in Neural Information Processing Systems 4*, pp. 993–1000. Morgan Kaufmann.

Obradovic, D. and Deco, G. (1995). Linear Feature Extraction in non-Gaussian Networks. In *World Congress on Neural Networks*, Vol. 1, pp. 523–526 Washington.

Parra, L. C., Deco, G., and Miesbach, S. (1995). Redundancy reduction with information-preserving maps. *Network: Computation in Neural Systems*, *6*, 61–72.

Pearlmutter, B. A. (1992). Temporally Continuous vs. Clocked Networks. In *Neural Networks in Robotics*, pp. 237–252. Kluwer Academic Publishers.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *Annals of Mathematical Statistics*, *22*, 400–407.

Torkkola, K. (1996a). Blind separation of convolved sources based on information maximization. In *Neural Networks for Signal Processing VI* Kyoto, Japan. IEEE Press. In press.

Torkkola, K. (1996b). Blind separation of delayed sources based on information maximization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* Atlanta, GA. In press.

Verhulst, P. F. (1844) *Nouveaux memoires de l'Academie royale des sciences et belles-lettres de Bruxelles*, *18*, 1. Also 1846, *20*, 1.

A Linear predictive source distributions

In the simulations of section 4 the $f_j(u_j; \mathbf{w}_j)$ distribution used is a mixture of logistic densities,

$$f_j(u_j(t)|u_j(t-1), u_j(t-2), \ldots; \mathbf{w}_j) = \sum_k m_{jk}\; h((u_j(t) - \bar{u}_{jk})/\sigma_{jk})/\sigma_{jk} \tag{5}$$

where σ_{jk} is a scale parameter for logistic density k of source j and is an element of $\mathbf{w}_j$, and the mixing coefficients m_{jk} are elements of $\mathbf{w}_j$ and are constrained by $\sum_k m_{jk} = 1$. The component means $\bar{u}_{jk}$ are taken to be linear functions of the recent values of that source,

$$\bar{u}_{jk} = \sum_{\tau=1} a_{jk}(\tau)\; u_j(t-\tau) + b_{jk} \tag{6}$$

where the linear prediction coefficients $a_{jk}(\tau)$ and bias b_{jk} are elements of $\mathbf{w}_j$.

To perform stochastic gradient descent it is necessary to calculate the derivative $df_j(u_j; \mathbf{w}_j)/d\mathbf{w}_j$. We accomplish this using the following equations. For conciseness, when we below refer to f_j, h_{jk}, and their simple derivatives f'_j, h'_{jk}, we leave off the arguments, which are the same as the corresponding arguments above. The h logistic density function and its cumulative distribution function g are as in footnote 1.

$$\frac{d\widehat{G}}{dm_{jk}} = -\frac{h_{jk}}{\sigma_{jk}\; f_j} \tag{7}$$

$$h'_{jk} = h_{jk}(1-2g) \tag{8}$$

$$\frac{d\widehat{G}}{d\sigma_{jk}} = \frac{(h_{jk}\;\sigma_{jk} + (u_j - \bar{u}_{jk})h'_{jk})m_{jk}}{\sigma^3_{jk}\; f_j} \tag{9}$$

$$\frac{d\widehat{G}}{da_{jk}(\tau)} = \frac{m_{jk}\; h'_{jk}\; u_j(t-\tau)}{\sigma^2_{jk}\; f_j} \tag{10}$$

$$\frac{d\widehat{G}}{db_{jk}} = \frac{m_{jk}\; h'_{jk}}{\sigma^2_{jk}\; f_j} \tag{11}$$

$$f'_j = \sum_k \frac{m_{jk}\; h'_{jk}}{\sigma^2_{jk}} \tag{12}$$

After each weight update the mixing coefficients must be normalized, $m_{jk} \leftarrow m_{jk}/\sum_{k'} m_{jk'}$.

B Stochastic gradient descent

In the above experiments a number of techniques were used to improve the efficiency and robustness of the stochastic gradient descent procedure as applied to cICA.

First, rather than performing gradient descent directly on the scale parameters σ_{jk} and mixing parameters m_{jk}, we performed gradient descent upon their logarithms. Using such log scale parameters automatically guarantees $\sigma_{jk} > 0$. In addition, the stability and robustness of the gradient descent process are improved (Nowlan and Hinton 1992; Pearlmutter 1992).

Second, an important contribution to the computational efficiency of our experiments is due to Amari, Cichocki, and Yang (1996), who post-multiply their ICA-like gradient by $\mathbf{W}^T\mathbf{W}$. Since this is a positive-definite matrix it does not effect the stochastic gradient convergence criteria, and the resulting quantity

$$\Delta\mathbf{W} \propto -\frac{d\widehat{G}}{d\mathbf{W}}\mathbf{W}^T\mathbf{W} = \mathbf{W} + \left(\frac{f'_j(u_j; \mathbf{w}_j)}{f_j(u_j; \mathbf{w}_j)}\right)_j \mathbf{u}^T\mathbf{W} \tag{13}$$

is therefore an admissible pseudo-gradient. This post-multiplication neatly eliminates the matrix inversion, and makes the algorithm scale-invariant to the true mixing matrix $\mathbf{A}$.