Proceedings 1995 International
Symposium on Non-linear Theory
and Applications 1:43-47.
(In press)

Fast blind separation based on information theory.

Anthony J. Bell & Terrence J. Sejnowski

Computational Neurobiology Laboratory,
The Salk Institute,
PO Box 85800,
San Diego, CA 9186-5800

ABSTRACT

Blind separation is an information theoretic problem, and we have proposed an information theoretic 'sigmoid-based' solution [2]. Here we elaborate on several aspects of that solution. Firstly, we argue that the separation matrix may be exactly found by maximising the joint entropy of the random vector resulting from a linear transformation of the mixtures followed by sigmoidal non-linearities which are the cumulative density functions of the 'unknown' sources. Secondly, we present the learning rule for performing this maximisation. Thirdly, we discuss the role of prior knowledge of the c.d.f.'s of the sources in customising the learning rule. We argue that sigmoid-based methods are better able to make use of this prior knowledge than cumulant-based methods, because the optimal non-linearity they should use is just an estimate of the source c.d.f. We also suggest that they may have the edge in terms of robustness and speed of convergence. Improvements in convergence speed have been facilitated by the introduction of pre-whitening of the mixture data. An example result demonstrating this is the perfect separation of ten artificially mixed audio signals in 10 seconds of workstation computing time (4 to prewhiten and 6 to separate).

I. Blind signal processing

Statistically independent sources propagating in a medium are subject to several forms of distortion and interference. They may be (1) mixed with other sources (2) mixed with time delayed versions of themselves, and (3) time-delayed. The mixing may be linear or non-linear. The inversion of these three forms of scrambling without any knowledge of their form may be called *blind signal processing*, or *blind identification*. When the mixing is linear, we usually refer to (1) as the problem of *blind separation* [4], (2) as the problem of *blind deconvolution*, and (3) as the problem of *blind time alignment*.

These problems are *information theoretic* problems in the sense that we are dealing with the removal of statistical dependencies introduced by the medium, and the correct measure of statistical dependency is *mutual information* (see below). In the most general information theoretic formalism, no special status is given to *noise* introduced by the medium or the sensors. It is regarded as another 'source' to be separated out. It cannot be assumed to be characterised only by second-order statistics (gaussian). In fact, if we are lucky (and we usually are), it will not be gaussian, for it is the higher-order statistics which characterise a signal as independent and enable it to be separated out from others.

In [2], an information theoretic approach was outlined to all three of the above problems. This paper is really a series of footnotes to [2], and should be read in conjunction with it if fuller details, or material of an introductory or tutorial nature are needed. Here we will concentrate on the blind separation problem in order to show more clearly how it is solved by information theory.

II. Separation through information theory.

A vector of sources $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]$ propagates in a medium and mixtures of them, $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)] = \mathbf{A}\mathbf{s}(t)$[1], are picked up by sensors. The mixing is linear and static, there are no time delays and there are the same number (N) of sensors as sources so that the mixing matrix, $\mathbf{A}$, is square.

The important fact that distinguishes a source, s_i, from a mixture, x_i, is that it is statistically indepen-

[1] henceforth, for convenience, the time index will be considered as implicit.

dent from the other sources, s_j. Their joint probability density function (p.d.f.), measured across the time ensemble, factorises:

$$f_{\mathbf{S}}(\mathbf{s}) = \prod_{i=1}^{N} f_{s_i}(s_i) \tag{1}$$

Another way of saying this is that the mutual information between any two sources, i and j, is zero:

$$I(s_i, s_j) = E\left[\ln \frac{f_{\mathbf{S}}(\mathbf{s})}{\prod_{i=1}^{N} f_{s_i}(s_i)}\right] = 0 \tag{2}$$

where $E[.]$ denotes expected value across the time ensemble. Mixtures of sources will be statistically dependent on each other and the mutual information between them, $I(x_i, x_j)$ will in general be positive. Blind separation then consists in finding a matrix, $\mathbf{W}$, so that the linear transformation $\mathbf{u} = \mathbf{Wx} = \mathbf{WAs}$ reestablishes the condition $I(u_i, u_j) = 0$, for all $i \neq j$. This is the problem of Independent Component Analysis (ICA) [4, 3] One solution to this problem is that $\mathbf{W}$ is the inverse of $\mathbf{A}$ so that $\mathbf{WA}$=$\mathbf{I}$, the identity matrix. Any other solution matrix, $\mathbf{W}$, can be shown to be a permutation and rescaling of this one. See Comon [3] for a fuller discussion of these matters.

To make the u_i independent, we need to operate on non-linearly transformed output variables, $y_i = g(u_i)$, $g()$ being a *sigmoidal* function.[2] The sigmoidal function provides, through its Taylor series expansion, all the higher-order statistics necessary to establish independence. This assertion is justified through the following theorem:

Theorem. Independent Component Analysis (blind separation) can be performed *exactly*, by finding the maximum, with respect to $\mathbf{W}$, of the joint entropy, $H(\mathbf{y})$, of an output vector, $\mathbf{y}$, which is the vector $\mathbf{u}$, except that each element is transformed by a sigmoidal function which is a c.d.f. of a sources which we are looking for.

In practice, we will often assume that all the sources have the same c.d.f. and use the same sigmoidal function for each element of $\mathbf{u}$. To prove this theorem, we develop the following six points:

Point 1. Independent variables cannot become dependent by passing each one through a sigmoid. Thus if $I(u_i, u_j) = 0$ and $\mathbf{y}$=$g(\mathbf{u})$, $g()$ being invertible, then $I(y_i, y_j) = 0$. Since g^{-1} is also invertible, the converse also holds.

[2] a sigmoidal function is defined somewhat generally here as an invertible twice-differentiable function mapping the real line into some interval, often the unit interval: $\mathbf{R} \rightarrow [0, 1]$.

Point 2. The entropy, $H(y)$, of a sigmoidally transformed variable has its maximum value (of zero) when the sigmoid function is the cumulative density function (c.d.f.) of the u-variable. Proof: $H(y)$ is maximum when $f_y(y)$=1 (the uniform distribution). Thus by the relation:

$$f_y(y) = \frac{f_u(u)}{dy/du} \tag{3}$$

we have $dy/du = f_u(u)$ which means $y = F_u(u)$, the cumulative density.

Point 3. The joint entropy, $H(y_1, y_2)$, of two sigmoidally transformed variables has its maximum value (of zero) when y_1 and y_2 are independent and the sigmoid function in each is the c.d.f. of u_1 and u_2 respectively. This is a clear consequence of Point 2 and the relation:

$$H(y_1, y_2) = H(y_1) + H(y_2) - I(y_1, y_2) \tag{4}$$

The N-variable joint entropy, $H(\mathbf{y})$, is similarly maximal when each $f_{y_i}(y_i)$ term is maximum and all the $I(y_i, y_j)$ are zero.

Point 4. When two independent non-gaussian variables, u_i and u_j are linearly combined, the p.d.f. of the resulting variable has a different shape from either of $f_{u_i}(u_i)$ or $f_{u_j}(u_j)$. In general, the p.d.f. becomes *more gaussian*, a trend ultimately enshrined in the Central Limit Theorem. Gaussian variables are the only ones which retain the form of their p.d.f. under linear combination.

Point 5. Consider the joint entropy, $H(\mathbf{y})$, of N sigmoidally transformed variables, where the sigmoid functions are the c.d.f.'s of N independent non-gaussian sources (ie: $y_i = F_{s_i}(u_i)$). This has its maximal value when $u_i = s_i$, in other words when the sources are separated! Any mixing of sources, $u_i = \sum_j s_j$, will both:

- introduce statistical dependencies between the u's, moving $I(u_i, u_j)$ away from zero (and hence also $I(y_i, y_j)$ — see Point 1), and
- decrease the individual entropy terms, $H(y_i)$, through deviation of $f_{y_i}(y_i)$ from 1.

This latter fact is born out by Points 2 and 4 above. Taken together, this shows that under the special condition that $y_i = F_{s_i}(u_i)$, the joint entropy $H(\mathbf{y})$ is maximal when the individual entropies, $H(y_i)$, are maximal and the mutual informations, $I(y_i, y_j)$ are minimal, conditions only satisfied by the separation solution, $u_i = s_i$.

Point 6. Therefore we can do blind separation by maximising the joint entropy, $H(\mathbf{y})$, of an output which has been transformed by sigmoids which are the c.d.f.'s of the sources we are looking for.

Maximisation of $H(\mathbf{y})$ is not difficult using standard stochastic gradient techniques common in neural networks work and elsewhere. Here we shall give a terse presentation. Full details and a more intuitive account are given in Bell & Sejnowski 1994. We utilise the multivariate version of Eq.3 [7]:

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{f_{\mathbf{x}}(\mathbf{x})}{|J|} \tag{5}$$

where $|J|$ denotes the absolute value of the determinant of the jacobian matrix:

$$J = \det \left[\frac{\partial y_i}{\partial x_j} \right]_{ij} \tag{6}$$

This relation enables us to write the joint entropy as:

$$H(\mathbf{y}) = -E[\ln f_{\mathbf{y}}(\mathbf{y})] = E[\ln |J|] + H(\mathbf{x}) \tag{7}$$

In gradient ascent, we change our $\mathbf{W}$ matrix over time proportional to the entropy gradient:

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} = E\left[\frac{\partial \ln |J|}{\partial \mathbf{W}} \right] \tag{8}$$

In *stochastic* gradient we remove the expected value operator and the derived rule is:

$$\Delta \mathbf{W} \propto [\mathbf{W}^T]^{-1} + \hat{\mathbf{y}} \mathbf{x}^T \tag{9}$$

where $\hat{\mathbf{y}} = [\hat{y}_1 \ldots \hat{y}_N]^T$, the elements of which are:

$$\hat{y}_i = \frac{\partial}{\partial y_i} \frac{\partial y_i}{\partial u_i} \tag{10}$$

When the sigmoids are the source c.d.f.'s $[y_i = F_{s_i}(u_i)]$ then this has the interesting form:

$$\hat{y}_i = \frac{\partial f_{s_i}(u_i)}{\partial F_{s_i}(u_i)} \tag{11}$$

Often, however, we will use a standard sigmoid function. For example, for the 'logistic' function, $y = (1 + \exp(-u))^{-1}$, we derive $\hat{y} = 1 - 2y$, and for the hyperbolic tangent function, $y = \tanh(u)$, we derive $\hat{y} = -2y$.

If the mixtures are not zero-mean, then it may be desirable to simultaneously train a vector of bias weights, $\mathbf{w}$, (so that $\mathbf{u} = \mathbf{W}\mathbf{x} + \mathbf{w}$). The rule for this is:

$$\Delta \mathbf{w} \propto \hat{\mathbf{y}} \tag{12}$$

The learning rule Eq.9 converges, meaning $\langle \Delta \mathbf{W} \rangle = \mathbf{0}$, when $\mathbf{I} = -\langle \hat{\mathbf{y}} \mathbf{u}^T \rangle$. The off-diagonal elements must be zero, and in general they expand to form a condition involving an infinite number of higher-order statistics. In the case which we argued above leads to independence, $[y_i = F_{s_i}(u_i)]$, this condition is, for $i \neq j$:

$$\left\langle \frac{\partial f_{s_i}(u_i)}{\partial F_{s_i}(u_i)} u_j \right\rangle = 0 \tag{13}$$

These results in this section may be derived within other superficially different formalisms (maximum likelihood, Kullback-Liebler distances etc) without altering their essential content.

III. Performance.

A. *Pre-whitening and Convergence Speed.*

Convergence speed using just the algorithm of Eq.9 and Eq.12 can be very slow, taking many hours to separate 10 signals, as reported in [2]. However, if we pre-process our training data to remove first and second order statistics, the speedups can be enormous. This process, called *pre-whitening*, or *sphering*, subtracts the means and decorrelates the inputs, giving each unit variance. The use of this method in conjunction with blind separation methods very similar to ours has been pioneered by Karhunen et al [5].[3]

The speedups achievable by prewhitening make possible the processing of very high dimensional data in reasonable time. For example, we are now performing experiments on natural images with input arrays as large as N=256. Convergence times on such data are comparable to those which, without prewhitening, we experienced for N=10.

In pre-whitening, after mean-subtraction, we multiply our data by a matrix $\mathbf{V}$ to make the covariance matrix of our data into the identity matrix:

$$\mathbf{x} \leftarrow \mathbf{V}(\mathbf{x} - \langle \mathbf{x} \rangle) \quad \text{after which} \quad \langle \mathbf{x}\mathbf{x}^T \rangle = \mathbf{I} \tag{14}$$

But which $\mathbf{V}$ should we choose, since there are many ways to decorrelate? Principal Component Analysis (PCA) is one way, choosing axes according to the directions of greatest variance in the data. We would like, however, to decorrelate in a way which makes subsequent ICA training speedier. One method which seems to works well in our simulations is to set $\mathbf{V}$ as follows:

$$\mathbf{V} = 2\sqrt{\langle \mathbf{x}\mathbf{x}^T \rangle^{-1}} \tag{15}$$

which actually makes $\langle \mathbf{x}\mathbf{x}^T \rangle = 4\mathbf{I}$. The interesting thing about this solution is that is exactly a scaled version (by a factor of $2\sqrt{2}$) of that which can be derived analytically by maximising entropy through a layer of outputs passed through the erf (or cumulative gaussian) non-linearity.[4] If $\mathbf{y} = \mathrm{erf}(\mathbf{u}) = \mathrm{erf}(\mathbf{W}\mathbf{x})$, then

[3] We are very grateful to Kari Torkkola for drawing their results to our attention.

[4] This fact was utilised by Baram & Roth [1], who proposed using the erf solution as a weight initialisation scheme for training these networks with the tanh function.

Eq.10 evaluates as $\hat{\mathbf{y}} = -2\mathbf{u}$ (see [2]) and the learning rule, as in Eq.9, evaluates to:

$$\Delta \mathbf{W} \propto [\mathbf{W}^T]^{-1} - 2\mathbf{u}\mathbf{x}^T \tag{16}$$

This stabilises, at $\mathbf{W} = (2\langle \mathbf{x}\mathbf{x}^T \rangle)^{-1/2}$ when $2\langle \mathbf{u}\mathbf{u}^T \rangle = \mathbf{I}$, showing that it has exactly the property we required in Eq.14, as a decorrelated input to further entropy-maximisation network.

We might call the stable solution to Eq.16, the 'Gaussian Component Analysis' solution (or GCA, to distinguish it from ICA and PCA), since it is the decorrelated solution which gives the most gaussianly distributed outputs.

B. Results.

The power of these methods is demonstrated by the fact that we have been able to separate ten audio signals to an average signal-to-noise ratio of 26dB in *two passes* through the data, one to prewhiten the input and one to separate. This result was achieved as follows (more details in [2]).

Ten speech or music samples, sampled at 8kHz and lasting six seconds were mixed together by a matrix of random values between 0 and 1. The time index of the data was permuted to make the sample stationary in time. The mean vector and covariance matrix of the data were computed and the data was then prewhitened using Eq.15. The $\mathbf{W}$ matrix was initialised to the identity matrix and then trained using Eq.9 with the logistic sigmoid non-linearity. Because of the relative computational expense of the $\mathbf{W}$-inversion, $\mathbf{W}$ was updated in 'batch' mode, meaning the $\Delta\mathbf{W}$'s were accumulated, in this case over 30 data vector presentations, and then the weight update took place. The learning rate (the proportionality constant in Eq.9) was 0.001. The simulations were performed using efficient vectorised MATLAB code on a Sparc-20 workstation.

Separation of these signals was effectively a real-time process, taking 4 seconds for the prewhitening pass and 6 seconds for the single separation pass. Further passes through the data, with an annealed (reduced) learning rate, were able to raise the signal-to-noise ratio to 36dB.

IV. Discussion.

A. Prior information and approximation.

Separation methods, although they seem to be completely unsupervised, do actually involve prior assumptions — they are model-dependent. In principle, to separate independent signals requires an infinite amount of prior information. This can be seen in the factorisation condition alone, which is a condition on functions, not variables:

$$f_{\mathbf{u}}(\mathbf{u}) = \prod_{i=1}^{N} f_{u_i}(u_i). \tag{17}$$

There are two main ways of parameterising these functions. In cumulant-based methods [3], to fully represent the above statistical condition we would need to evaluate all higher-order moments up to infinity, an impossible task. In sigmoid-based methods based on our entropy maximisation algorithm, we would need to know the exact form of the p.d.f.'s of the sources. These forms of knowledge amount to the same thing (a p.d.f. can be transformed to a characteristic function, which provides the moments [7]). Since in general, our knowledge will be incomplete, the question becomes which basis (cumulants or p.d.f.'s) will we choose in order to perform our approximations?

Before arguing for p.d.f.'s over cumulants, it is worth noting what approximations are typically made in practice. In the case of cumulant methods, usually moments are evaluated only up to fourth order [3]. In the case of sigmoid methods, we usually use the 'logistic' sigmoid, $y = (1 + \exp(-u))^{-1}$, or, for equivalent effect, the tanh function [2]. Clearly, both approximations may produce networks that fail to separate certain signals. It is an interesting and difficult theoretical problem to show under what conditions this may occur. Based on analyses by Moreau & Macchi [6] and on empirical results in [2], the current understanding seems to be that both methods are successful when the p.d.f.'s of the sources are super-gaussian — when their kurtosis is greater than zero. However, more work has to be done to understand these issues.

B. Advantages of sigmoid methods.

Nonetheless, we believe that there are several advantages that sigmoid-based methods have over cumulant-based methods.

1. **Performance.** Incorporating the pre-whitening techniques described in Section 3.2, sigmoid-based methods are capable of separating 10 audio signals in about 6 seconds on a workstation. Furthermore, they have not yet been observed to converge to an incorrect solution on digitally mixed data. Convergence speed is now high enough that we have been able to start dealing (in other domains) with very high dimensional data sets with on the order of several hundred inputs.

2. **Sigmoid customisation.** Sigmoid methods approximate the source p.d.f. *before* deriving the information theoretic rule of Eq.9, whereas cumulant methods expand the mutual information, and *then* truncate the expansion. In

choosing the former path ('approximate then derive'), we are able to customise our sigmoid function before deriving the exact form of the $\hat{y}_i$ terms in Eq.10, and thus preserve *all* higher-order statistics in the process. This is a much more attractive course of action than trying to decide which higher-order cumulants must be given more weight in a p.d.f.-dependent cumulant method. In sigmoid methods, we can give the sigmoid function any shape as long as it is bounded and monotonic, and each has a very natural interpretation as a (possibly scaled and shifted) c.d.f. of the source which we are attempting to separate. In many cases, we will be able to directly sample this c.d.f. in some pre-training phase, as when we measure a typical signal in the absence of interference. Then we can build a lookup table based on the sampled c.d.f. and use it as our sigmoid function. In this case, for the operation of the algorithm in Eq.8, the only thing that actually has to be stored is the lookup table values of $\hat{y}$, which must be calculated from the sampled c.d.f. and Eq.10.

In practice, it will very often not be necessary to use lookup tables because of the next point.

3. **Robustness.** All simulations we have performed on audio signals have converged correctly despite the fact that the logistic sigmoid (or the hyperbolic tangent) function which we use, is not a good fit for the c.d.f.'s of audio signals, which are typically more kurtotic [2], (or 'super-gaussian'). In this case we cannot make the arguments of Section 2, about the $H(y_i)$ terms having the same maxima as the minima of the $I(y_i, y_j)$ terms. In fact, it is clear that we may achieve higher $H(y_i)$ terms with linear combinations of super-gaussian sources, since they will better 'fit' the gradient of the logistic function. Nonetheless, (see Eq.4), the mutual information introduced by such combinations is apparently great enough, in the case of audio signals, to make such combinations disadvantageous for the overall entropy maximisation. The limits of this robustness must, of course, be more assiduously tested.

In conclusion, we have demonstrated rigorously how an information theoretic approach solves the problem of blind separation, as well as showing how convergence may be accelerated greatly. We believe that these factors, together with the potential for customising the algorithm for different data, make our approach an attractive one.

Acknowledgement

The authors are both with the Howard Hughes Medical Institute in the Computational Neurobiology Laboratory of the Salk Institute. This work was funded by a grant from the Office of Naval Research, and by the Howard Hughes Medical Institute. We are grateful for many helpful discussions with Paul Viola and Nicol Schraudolph.

References

[1] Baram Y. & Roth Z. 1994. Multi-dimensional density shaping by sigmoidal networks with application to classification, estimation and forecasting, CIS report no. 9420, October 1994, Centre for Intelligent systems, Dept. of Computer Science, Technion, Israel Inst. of Technology, Haifa, submitted for publication

[2] Bell A.J. & Sejnowski T.J. (1995). An information maximization approach to blind separation and blind deconvolution, *Neural Computation*, 7, 1129-1159 (also available from ftp SITE: ftp.salk.edu, FILE: pub/tony/bell.blind.ps.Z)

[3] Comon P. 1994. Independent component analysis, a new concept? *Signal Processing*, 36, 287-314

[4] Jutten C. & Herault J. 1991. Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture, *Signal processing* 24, 1-10

[5] Karhunen J., Wang L. & Jousensalo J. 1995. Neural estimation of basis vectors in Independent Component Analysis, *Proc. ICANN, Paris, 1995*

[6] Moreau E. & Macchi O. 1993. *Proc. IEEE Signal Processing Workshop on Higher Order Statistics*, Lake Tahoe, USA, June 1993, pp.215-219

[7] Papoulis A. 1984. *Probability, random variables and stochastic processes, 2nd edition*, McGraw-Hill, New York