BLIND SOURCE SEPARATION BY LOCAL INTERACTION OF OUTPUT SIGNALS

John W. Fisher III

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA 02139
fisher@ai.mit.edu

José C. Principe

Computational NeuroEngineering Laboratory
University of Florida
451 EB #33
Gainesville, FL 32611
principe@cnel.ufl.edu

ABSTRACT

We compare independent component analysis (ICA) [Bell and Sejnowski, 1995] to an alternative method [Fisher and Principe, 1996] for blind source separation of instantaneous linear mixtures. The method and its application to blind source separation of instantaneous linear mixtures is reviewed. Empirical results separating sources of varying kurtosis and a limited number of samples are presented. We demonstrate empirically that despite the additional computational cost of the method presented, significantly better performance can be achieved for a small number of samples and when the kurtosis of the sources is sub-gaussian. The method is also of interest as it can be extended easily to any differentiable nonlinear mapping.

1. INTRODUCTION

Blind source signal separation has garnered renewed attention in the signal processing community. A popular approach for the separation of instantaneous linear mixtures is the technique of Bell and Sejnowski [1] typically referred to as ICA (independent components analysis). The ICA approach is attractive because it results in a simple adaptive approach for estimating the separating function while placing relatively mild constraints on the input sources.

We have recently introduced an alternative method for the extraction of statistically independent features [3, 4, 5]. This paper discusses the differences in the two approaches and in particular compares their small sample performance as well as the effect of source kurtosis. Entropy maximization lies at the heart of both methods. While the ICA approach maximizes entropy by operating on the Jacobian of the mapping, our approach seeks to maximize entropy by operating on the observed outputs.

In this paper we discuss the application of this method to the problem of blind source separation of linearly mixed signals. We present empirical results comparing the effectiveness of both methods for the separation of sub-gaussian and super-gaussian sources. In so doing, we illustrate some of the relevant differences and similarities between the two approaches.

1.1 Blind Source Separation of Linearly Mixed Sources

Blind source separation of instantaneously and linearly mixed signals is described in a straightforward manner. A set of N unknown source signals, $s_i(t)$, are linearly and instantaneously mixed by an *unknown* mixing matrix $\boldsymbol{A}$, to yield $M \geq N$ mixed signals, $x_i(t)$. In matrix form the set of signals observed at time, t, is written

$$\left[x_1(t) \ \ldots \ x_M(t)\right]^{\mathrm{T}} = \boldsymbol{A}\left[s_1(t) \ \ldots \ s_N(t)\right]^{\mathrm{T}} ; \boldsymbol{A} \in \Re^{M \times N} \quad . \quad (1)$$

In blind source separation we seek to recover the original source signals from the mixture signals. It is possible to do so without knowledge of the mixing matrix when the following conditions are satisfied: [2]

1. The mixing matrix, $\boldsymbol{A}$, has full column rank.
2. The source signals $s_i(t)$ are mutually independent.
3. At least $N-1$ of the source signals are non-Gaussian.

If the above conditions hold, then any matrix $\boldsymbol{B} \in \Re^{N \times M}$, which results in random processes, $\hat{s}_i(t)$,

$$\begin{aligned}\left[\hat{s}_1(t) \ \ldots \ \hat{s}_N(t)\right]^{\mathrm{T}} &= \boldsymbol{B}\left[x_1(t) \ \ldots \ x_M(t)\right]^{\mathrm{T}} \\ &= \boldsymbol{BA}\left[s_1(t) \ \ldots \ s_N(t)\right]^{\mathrm{T}} \\ \hat{\boldsymbol{s}}(t) &= \boldsymbol{Bx}(t) = \boldsymbol{BAs}(t)\end{aligned}$$

such that the $\hat{s}_i(t)$ are mutually independent, will recover the original source signals up to a scaling and permutation. In other words, the matrix $\boldsymbol{B}$ along with the unknown mixing matrix $\boldsymbol{A}$ will satisfy the relation

$$\boldsymbol{BA} = \Lambda \boldsymbol{P} \boldsymbol{I}_N \, , \quad (2)$$

where Λ is a real diagonal matrix, $\boldsymbol{P}$ is a permutation matrix, and $\boldsymbol{I}_N$ is the N-dimensional identity matrix. In general, equation 2 is satisfied only approximately.

This work was supported by AFOSR MURI through Boston University GC123919NGD.

2. ICA AND LOCAL INTERACTION METHODS

Both methods, either directly or indirectly seek the coefficients of the matrix, $\boldsymbol{B}$, such that the elements of the vector, $\boldsymbol{y}$, which are the outputs of the mapping

$$\begin{aligned} \boldsymbol{y} &= g(\hat{\boldsymbol{s}}) \\ &= g(\boldsymbol{B}\boldsymbol{x}) \end{aligned} \tag{3}$$

approach statistical independence. In 3, $g(\boldsymbol{u})$ is the logistic function ($g(\boldsymbol{u}) = (1 + \exp(\boldsymbol{u}))^{-1}$), exploiting the property that statistical independence is maintained when signals are passed through a monotonic transformation.

If the elements of $\boldsymbol{y}$ are statistically independent and conditions 1 through 3 above are met, then the elements of $\hat{\mathbf{s}}$ will have, in some measure, recovered the original sources. From an algorithmic viewpoint, any differentiable, monotonic and saturating nonlinearity is suitable for either approach; however, strictly speaking the nonlinearity should match the cumulative distribution functions (CDF) of the sources in order to achieve maximum entropy [2]. In practice, it has been shown that exact matching of the CDF is not necessary for blind source separation [1].

The ICA approach maximizes the differential entropy of the mapping directly [1]. Consequently the adaptation of the de-mixing matrix becomes

$$\Delta\boldsymbol{B} \propto \boldsymbol{B}^{-\mathrm{T}} + (1 - 2\boldsymbol{y})\boldsymbol{x}^{\mathrm{T}} .$$

As increasing differential entropy increases statistical independence, the outputs, $\hat{\boldsymbol{s}}$, should recover the original sources, $\boldsymbol{s}$.

In contrast, the local interaction method manipulates entropy from the output samples. The measure of statistical independence is the integrated squared error (ISE) between the *estimated* density, $\hat{f}_{\boldsymbol{Y}}(\boldsymbol{u}, \{\boldsymbol{y}\})$, over samples of $\boldsymbol{y}$ at the output of the mapping and the uniform density, $f_{\boldsymbol{U}}(\boldsymbol{u})$ [3, 4],

$$\begin{aligned} \mathbf{ISE} &= \frac{1}{2}\int_{\Omega_{\boldsymbol{Y}}} \left(f_{\boldsymbol{U}}(\boldsymbol{u}) - \hat{f}_{\boldsymbol{Y}}(\boldsymbol{u}, \{\boldsymbol{y}\})\right)^2 d\boldsymbol{u} \\ &= \frac{1}{2}\int_{\Omega_{\boldsymbol{Y}}} \left(f_{\boldsymbol{U}}(\boldsymbol{u}) - \hat{f}_{\boldsymbol{Y}}(\boldsymbol{u}, \{g(\boldsymbol{B}x)\})\right)^2 d\boldsymbol{u} \end{aligned} \tag{4}$$

exploiting the property that entropy is maximized for the uniform density when the mapping is restricted to a finite range. It can be shown that the ISE criterion is equivalent to expanding differential entropy as a Taylor series up to second order. In equation 4 we employ the Parzen window density estimator using Gaussian kernels, although other nonparametric density estimators are suitable. Minimization of the ISE criterion implies estimation of the output probability density function, but such is not the case. We have shown that the adaptation resulting from the ISE criterion can be computed simply and exactly by evaluating the expression

$$\begin{aligned} \Delta\boldsymbol{B} &\propto \sum_i ([\boldsymbol{y}_i \bullet (\mathbf{1} - \boldsymbol{y}_i)]\varepsilon_i^{\mathrm{T}}) \\ \varepsilon_i &= f_r(\boldsymbol{y}_i) - \sum_{\neq} \kappa_a(\boldsymbol{y}_i - \boldsymbol{y}_j) \end{aligned} \tag{5}$$

at the locations of the signal samples in the output space[5]. The notation $\bullet$ indicates element by element vector multiplication, $\kappa_a(\ )$ is the local influence exerted by each data point in the output space, and $f_r(\ )$ is a boundary influence function.

The vector-valued local influence function, the convolution of the estimator kernel with its gradient, has the form

$$\kappa_a(\boldsymbol{y}) = -\left(\frac{1}{2^{N+1}\pi^{N/2}\sigma^{N+2}}\right)\exp\left(-\frac{1}{4\sigma^2}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})\right)\boldsymbol{y} \tag{6}$$

when the estimator kernel is an N-dimensional gaussian. Centered at each data point in the output space, this function is the influence each data point exerts on its locale.

The vector-valued boundary influence function, the convolution of the kernel with the desired output density, has the form

$$f_r(\boldsymbol{y}) = \begin{vmatrix} \frac{1}{a^N}\left(\prod_{i\neq 1}\frac{1}{2}\left(\mathrm{erf}\left(\frac{y_i}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{y_i - a}{\sqrt{2}\sigma}\right)\right)\right) \\ \times(\kappa_1(y_1, \sigma) - \kappa_1(y_1 - a, \sigma)) \\ \vdots \\ \frac{1}{a^N}\prod_{i\neq N}\frac{1}{2}\left(\mathrm{erf}\left(\frac{y_i}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{y_i - a}{\sqrt{2}\sigma}\right)\right) \\ \times(\kappa_1(y_N, \sigma) - \kappa_1(y_N - a, \sigma)) \end{vmatrix} \tag{7}$$

when the desired output density is uniform over an N-dimensional hypercube with vertices of size a ($\kappa_1(u, \sigma)$ is the one-dimensional gaussian kernel with standard deviation σ). This function is the influence that the boundaries of the uniform density exert on points near the boundary in the output space. In practice, the erf( ) terms can be ignored when the saturating nonlinearity is used [5].

Both the local and boundary influence functions are shown in figure 1 (not to scale). Intuitively the local influence function has a diffusive effect on output sample points while the border influence term prevents saturation.

3. EXPERIMENTAL RESULTS

In our experiments we compare the local interaction and ICA methods for separating instantaneous linear mixtures of sub-gaussian and super-gaussian sources. This set of experiments illustrates the comparative sensitivity of both methods to the kurtosis of the source distributions and small sample size.

There are three experiments:

1. separation of two sources with the same distribution,
2. separation of three sources with the same distribution,
3. separation of three sources with three different distributions.

In each experiment we conduct 20 monte carlo runs using the same mixing matrices.

The nominally well conditioned mixing matrices are either

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad \text{or} \qquad \boldsymbol{A} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

for two and three independent sources, respectively. In each trial the initial estimate of the de-mixing matrix is the identity matrix. As all sources have unit variance the initial signal-to-noise ratios are 6 and 3 dB in the two and three source experiments, respectively.

The sources are various combinations of three distributions; the laplace, logistic and uniform densities with relative kurtoses of 3, 1.2, and -1.2, respectively. The logistic density has a logistic CDF which is the nonlinearity used for these experiments (i.e. maximum entropy can be achieved for these sources). Figure 2 shows the densities and their associated distributions.

In every trial the local interaction and ICA methods use the same sample draw allowing direct trial comparisons. Each method was iterated (in batch) until the coefficients of the separating matrix estimate had converged (in the sense that its Frobenius norm did not change significantly). The number of samples (per source) in each trial is 100.

In the first set of experiments samples were drawn from two sources using the same distributions. This was repeated for each of the distribution types. Table 1 summarizes the results, showing when the local interaction approach performs better and worse than ICA. “SNR better in 2 sources” is the number of trials (out of 20) in which the local interaction method recovered both sources with a higher SNR than ICA. Likewise, “SNR worse in 2 sources” is the number of trials in which ICA similarly performed better than the local interaction approach. “min SNR (dB)” is the lowest SNR for either source for the local interaction method over all trials (with the same measure for ICA in parentheses - (NM) indicates that the ICA did not recover both sources in every trial with greater than 0.25 dB SNR). The “best difference” entry is the best SNR difference for a single recovered source over all trials for the local interaction approach versus ICA (VL indicates a trial in which ICA failed to recover either source with greater than 0.25 dB SNR). The “worst difference” entry is the opposite; the degree by which SNR of the ICA recovered source exceeded the SNR of the local interaction method (NM indicates that there was no trial in which ICA recovered either source with greater than 0.25 dB SNR).

In the laplace case (higher kurtosis), the two methods are comparable with the ICA approach performing slightly better (e.g. in 7 trials ICA recovered both sources with higher SNR versus 5 for the local interaction method). In the other cases; however, the local interaction approach outperforms ICA in nearly all measures. Notably, the local interaction approach exhibits some degree of source separation for all three distribution types. It is, of course, not surprising that ICA failed on the uniform source densities as this was pointed out in the Bell and Sejnowski paper [1], but it is encouraging that the local interaction approach did not suffer the same shortcoming.

Table 1: Performance comparison for separation of two independent sources with like distributions.

	laplace	logistic	uniform
SNR better in 2 sources	5	16	20
SNR worse in 2 sources	7	1	0
min SNR (dB)	10 (11)	17 (3)	9 (NM)
best difference (dB)	22	52	VL
worst difference (dB)	-28	-24	NM

Table 2 summarizes the results for the same set of experiments, however, with three sources instead of two. As in the previous case, the local interaction method is noticeably better than ICA for the logistic and uniform sources. Furthermore, the performance on laplace sources is nearly identical.

Table 2: Performance comparison for separation of three independent sources with like distributions.

	laplace	logistic	uniform
SNR better in 3 sources	0	7	20
SNR worse in 3 sources	5	1	0
SNR better in 2 sources	10	14	20
SNR better in 2 sources	10	4	0
min SNR (dB)	9 (10)	11 (NM)	9 (NM)
best difference (dB)	21	VL	VL
worst difference (dB)	-19	-19	NM

The last experiment is of more interest; in this case the sources are of both sub- and super- gaussian distributions. In this set of trials, we use three sources, but drawing from each of the three distributions. As one of the sources is uniform, it is not surprising that the local interaction approach outperforms ICA. It is surprising, however, that in 14 of the 20 trials the local interaction method did a superior job in separating the laplace source (in light of the first two experiments in which the performance was at best comparable). One might conclude that the addition of a single sub-gaussian source has significantly hindered ICA, although it may

also be attributed to the small number of samples. Further experimentation is warranted.

Table 3: Performance comparison for separation of three independent sources with mixed distributions.

	mixed
SNR better in 3 sources	12
SNR worse in 3 sources	0
SNR better in 2 sources	20
SNR worse in 2 sources	0
laplace better	14
logistic better	18
uniform better	20
min SNR (dB)	10 (NM)
best difference (dB)	$\approx \infty$
worst difference (dB)	-8

4. COMMENTS

We have presented some empirical results examining the effect of small sample size and varying kurtosis for blind source separation of instantaneous linear mixtures. We have demonstrated empirically that the local interaction approach described has superior performance in these conditions. We note that this does not come at small cost. The computation of the update term is quadratic in the number of samples while for ICA it is linear. Despite this we are encouraged by these results as there are problems which are indeed sample size limited (e.g. radar imagery) and in which the sources have low relative kurtosis (e.g. EEG data). Furthermore, we also note that with a limited number of samples the local interaction approach never achieved less than 9 dB of separation for any source type and never failed to achieve some degree of separation for all sources.

Examination of equation 5 reveals that the update rule for the local interaction approach uses an error direction term, ε_i, and as such is directly extensible to differentiable nonlinear functions using error backpropagation[4,5]. This allows for more complex mapping structures; although it is an open question as to how such structures might be used.

5. REFERENCES

[1] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution", *Neural Computation*, vol. 7, pp. 1129-1159, 1995

[2] S. Bellini, "Bussgang techniques for blind deconvolution and equalization", in *Blind Deconvolution*, S. Haykin, ed., pp. 8-59, Prentice Hall, Englewood Cliffs, NJ., 1994

[3] J. Fisher and J. Principe, "Unsupervised learning for nonlinear synthetic discriminant functions", *Proc. SPIE, Optical Pattern Recognition VII*, D. Casasent and T. Chao, Eds., 1996, vol. 2752, pp. 2-13.

[4] J. Fisher and J. Principe, "Entropy manipulation of arbitrary nonlinear mappings", *Proc. of the IEEE Workshop Neural Networks for Signal Processing VII*, J. Principe, Ed., 1997, pp. 14-23.

[5] J. Fisher and J. Principe, "A methodology for information theoretic feature extraction", *Proc. of the IEEE International Joint Conference on Neural Networks*, A. Stuberud, Ed., 1998.

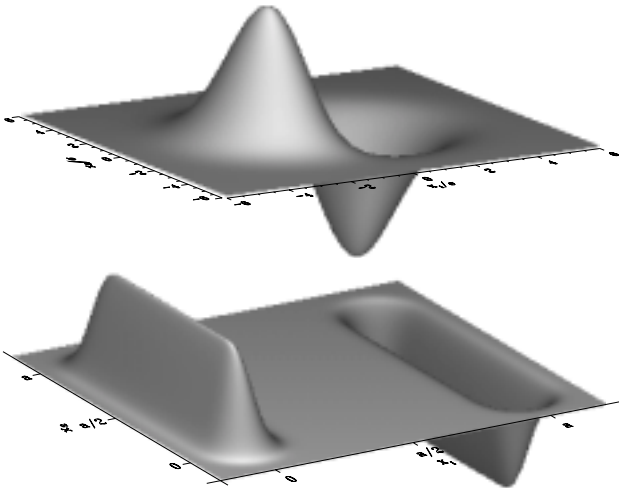

Figure 1: Local influence function (top) and boundary influence function (bottom) for $\mathbf{y}_1$-component. The $\mathbf{y}_2$-component is a 90 degree rotation counter-clockwise.

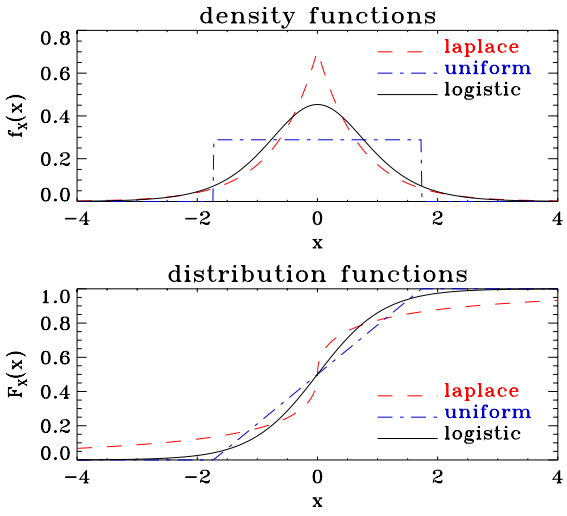

Figure 2: Density (top) and corresponding distribution (bottom) functions used for experiments.