INDEPENDENT COMPONENT ANALYSIS THROUGH DIRECT ESTIMATION OF THE MUTUAL INFORMATION

Georges A. Darbellay[1] and *Petr Tichavský*[2]

[1]Signal Processing Laboratory, EPFL/DE, Ecublens, CH-1015 Lausanne, Switzerland
e-mail: georges.darbellay@epfl.ch
[2]Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, P.O. Box 18, CZ-182 08 Prague, Czech Republic
e-mail: tichavsk@utia.cas.cz

ABSTRACT

We introduce an algorithm for the estimation of the mutual information between several signals. This algorithm also provides a convenient way of estimating the entropy of scalar signals. The algorithm uses a data-dependent partitioning of the observation space and has a relatively low computational cost. Matlab code is available. The usefulness of this approach in solving the independent component analysis problem is demonstrated on the separation of a linear mixture of two and three real-world speech data sets.

1. INTRODUCTION

Independent component analysis (ICA) may be understood as a generalization of principal component analysis or of factor analysis [1]. Instead of using covariances, one works with some measure of statistical dependence which takes into account the full dependence structure, or at least one which goes further than second order moments. The latter way consists in working with higher order moments. The former way is more ambitious and could be based on any one of the measures of (full) dependence which have been devised in statistics, or in related fields. In the area of blind signal separation [2, 3, 4, 5], the most popular such measure is undoubtedly the mutual information. Our contribution will aim at introducing into the problem of blindly separating mixed signals, a recently proposed estimator of the mutual information.

The mutual information $I(Y)$ between m scalar random variables Y_i, $Y = (Y_1, \ldots, Y_m)$, is defined as the Kullback-Leibler divergence between the joint density $f_Y(y)$ of Y and the product of its marginal densities $\prod_{i=1}^m f_{Y_i}(y_i)$,

$$I(Y) = \int f_Y(y) \log \frac{f_Y(y)}{\prod_{i=1}^m f_{Y_i}(y_i)} dy \tag{1}$$

This specific form of the Kullback-Leibler divergence is also called the redundancy. $I(Y) = 0$ if and only if $\{Y_i\}$ are independent, i.e. if $f_Y(y) = \prod_{i=1}^m f_{Y_i}(y_i)$ for almost all y. The mutual information may also be defined in terms of the differential entropy as $I(Y) = \sum_{i=1}^m h(Y_i) - h(Y)$, where

$$h(Y) = -\int f_Y(y) \log f_Y(y)\, dy. \tag{2}$$

The entropy is also important in ICA, because it often possible to work with orthogonal matrices only, in which case it is sufficient to minimize the sum of marginal entropies $\sum_{i=1}^m h(Y_i)$ (e.g. [3]).

Most ICA algorithms available in the literature avoid a direct estimation of the mutual information or the entropy. They resort to some approximations, usually of the entropy [9]. This may mean simplicity and speed. However, many of the known ICA algorithms make special requirements on the probability distributions of the signal sources.

Recently, an ICA algorithm called "FastICA", having very promising properties, was proposed [6]. The algorithm represents, in fact, a class of algorithms, because an arbitrary user-chosen nonlinear function can be selected. Depending on the signals to separate, one of these nonlinear functions is usually better than the others (as will become apparent in Tables 1 and 2). Thus, for this algorithm, as well as in general, it would be very desirable to have a measure of a residual mutual information between the separated signals. If the mutual information is not satisfactorily low, one can then iterate the procedure with a different type of nonlinearity.

In this contribution we demonstrate the suitability a recently published algorithm for estimating the mutual information. The estimator is presented in Section 2. Sections 3 and 4 contain numerical examples, with, respectively, two and three speech signals. Our conclusions are summarized in Section 5.

2. THE ESTIMATOR

In this section we will outline the algorithm. For more details we refer to [10]–[12]. For simplicity we focus on dimension two, i.e. on estimating the mutual information of two scalar random variables from data. Extension of the algorithm to higher dimensions is straightforward (but, obviously, the required number of observations significantly rises with the dimension).

Supported by the Grant Agency of the Academy of Sciences of the Czech Republic through grant K 1075601.

The algorithm is based on a partitioning of the observation space into a finite number of nonoverlapping rectangular cells C_k, $k = 1, \ldots, K$ as shown in Figure 1.

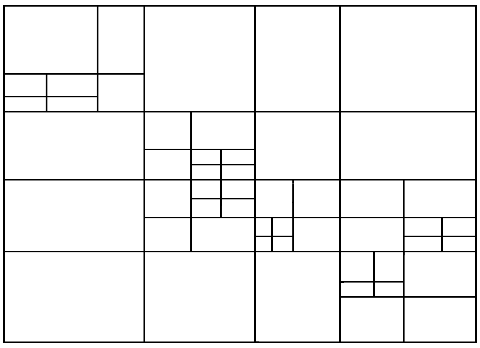

Fig. 1 Illustration of adaptive partitioning of the observation space.

Schematically, the algorithm may be formulated in the form of three rules.

(R0) Let the initial one-cell partition be the rectangle containing all data pairs (x,y).

(R1) Every cell is tentatively partitioned by dividing each one of its edges into 2 equiprobable halves (i.e. each one of the vertical and horizontal strips intersecting at the cell are divided into 2 halves, see Figure 2).

(R2) The tentative partitioning of a cell is accepted unless the ratio $f_{X,Y}(x,y)/[f_X(x)f_Y(y)]$ in (1) takes approximately the same value in each of its four subcells as it does in the cell itself.

The condition in (R2) means *conditional independence*. If, on every cell of a partition, conditional independence has been achieved, then it can be shown that

$$I(X,Y) = \sum_{k=1}^{K} P_{X,Y}(C_k) \log \frac{P_{X,Y}(C_k)}{P_X(C_k)P_Y(C_k)} \qquad (3)$$

where $P_{X,Y}(C_k)$ is the probability that the pair (X,Y) falls into the cell C_k, and $P_X(C_k)$,$P_Y(C_k)$ are the corresponding marginal probabilities. That is, if C_k is the product of intervals $C_k = (x_{k1}, x_{k2}) \times (y_{k1}, y_{k2})$, then $P_X(C_k)$ is the probability that $X \in (x_{k1}, x_{k2})$ and similarly $P_Y(C_k)$ denotes the probability that $Y \in (y_{k1}, y_{k2})$.

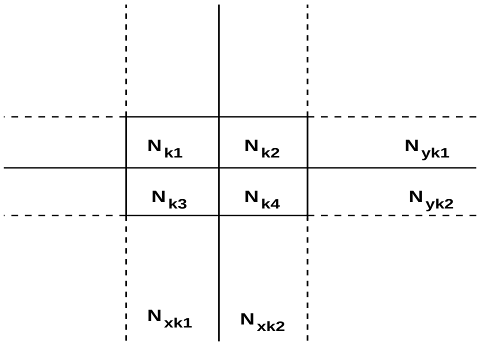

Fig. 2 Partitioning of one cell.

The partitioning proceeds recursively, as is apparent in Figure 1, and possesses a tree structure with 4 branches at each node. Let a rectangular cell contains at least $N_k > 4$ observations and let the corresponding vertical and horizontal strips contain, respectively, N_{xk} and N_{yk} observations, as shown in Figure 2. The cell is divided into 4 rectangular subcells containing $N_{k1}, \ldots, N_{k4}$ points so that the marginal numbers of points $N_{xk1} \geq N_{xk2}$ and $N_{yk1} \geq N_{yk2}$ differ as little as possible. If there are no repeating observations and if N_{xk} and N_{yk} are even, then $N_{xk1} = N_{xk2} = N_{xk}/2$ and $N_{yk1} = N_{yk2} = N_{yk}/2$.

The local independence would imply that $N_{k1} \approx N_{k2} \approx N_{k3} \approx N_{k4} \approx N_k/4$. We use the χ^2 "goodness-of-fit" test at the 5% significance level, that is if

$$\frac{4}{N_k} \sum_{i=1}^{4} \left(N_{ki} - \frac{N_k}{4}\right)^2 \leq \chi^2_{0.95}(3) = 7.81 \ . \qquad (4)$$

If the above condition is fulfilled, the hypothesis of conditional independence is accepted and the cell is not subjected to further partitioning (with the exception of the initial cell, whose partitioning is always performed). The cells containing $N_k < 4$ observations and those where the hypothesis of conditional independence was accepted, contribute to the estimated mutual information in (3) by the amount

$$\frac{N_k}{N} \log \frac{N_k/N}{N_{xk}/N \cdot N_{yk}/N} = \frac{N_k}{N} \log \frac{N_k}{N_{xk}N_{yk}} + \frac{N_k}{N} \log N$$

where N is the total number of observations. The matlab code of the described procedure in arbitrary dimension is available, [8].

Note that the algorithm above may be used to estimate not only the mutual information but also other measures of statistical dependence, such as the χ^2 or the Hellinger coefficient. It may happen that some measures are sharper (i.e. more sensitive) than others depending on the specific problem, but this question exceeds the scope of this paper.

The information being a difference of entropies, it is clear that the estimation of the mutual information may be reduced to the estimation of the entropy. In fact, the reverse is true as well. Let N be a random variable which is independent of X and suppose that we can easily calculate the entropy $h(N)$. For example, N could be a Gaussian variable. It is not difficult to show that

$$\begin{aligned} & I(X+N,N) - I(X+N,X) \\ & = h(X+N) - h(X+N|N) - h(X+N) + h(X+N|X) \\ & = h(N) - h(X). \qquad (5) \end{aligned}$$

Therefore

$$h(X) = h(N) - I(X+N,N) + I(X+N,X). \qquad (6)$$

Thus, in practice, the entropy may be estimated by generating a data sample for N and then calculating two mutual informations.

3. EXAMPLE 1: SEPARATION OF TWO SPEECH SIGNALS

The input data are two speech data files with American English words "black" and "white".[1] The data file were both shortened, i.e. we deleted a little at the beginning and at the end of the files so that both data sets were of equal length, $N = 5000$. Let the data be arranged in an $(N \times 2)$ matrix denoted $s = [s_1, s_2]$. We normalized the data so that both signals have a unit Euclidean norm, $\|s_1\| = \|s_2\| = 1$. Both signals are displayed in Figure 3.

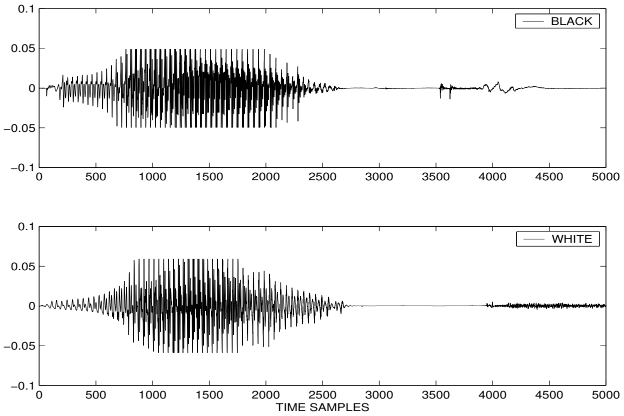

Fig. 3 Original speech signals.

The time-frequency structure of the signals is not taken into account for doing ICA: the data are just considered to be samples from probability distributions p_{s1} and p_{s2}. In this example, the mutual information between s_1 and s_2 was estimated by our algorithm to be $I(s_1, s_2) = 0.49$. The correlation coefficient of the data is 0.0371. The nonzero mutual information can be explained by the fact that both data files begin and end by periods of silence. Despite the lack of independence of the original signals we show that it is possible to recover the signals from an unknown linear mixture.

Assume that the observed data are given as

$$x = s \cdot M \tag{7}$$

where M is a 2×2 matrix, to be estimated. In our example we used

$$M = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix} . \tag{8}$$

The observed data (the mixtures) are plotted in Figure 4.

The first step of most of ICA procedures consists in removing the mean of the data and of *prewhitening* them. This means multiplying the data by a matrix W, $x' = xW$ so that the columns are uncorrelated, $x'^T x' = I$, where T denotes the transpose. This is done by choosing $W = (x^T x)^{-1/2}$.

The crucial step of each ICA procedure consists in finding a proper orthogonal (rotation) matrix that separates the prewhitened data to independent components. In $\mathbb{R}^2$, a general orthogonal matrix has the form

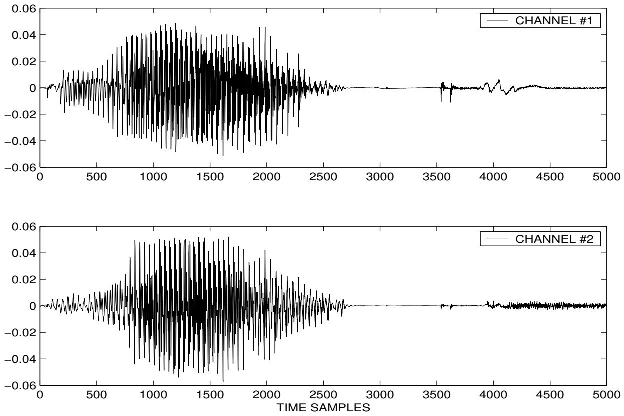

Fig. 4 Observed mixtures of the original signals.

$$Q(\alpha) = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \qquad \alpha \in (-\pi, \pi] . \tag{9}$$

With our algorithm, ICA can be performed by minimizing the mutual information of the columns of $xWQ(\alpha)$, respectively of $x'Q(\alpha)$, as shown in Figure 5. WQ may be understood as the demixing matrix. The mutual information as a function of α has four main minima in the interval $(-\pi, \pi]$, that all give the same result up to the order and sign of the separated components. The blindly recovered signal components are shown in Figure 6. The mixing matrix can be estimated by $\hat{M} = (WQ(\hat{\alpha}))^{-1}$, where $\hat{\alpha}$ minimizes the mutual information. In our case we got

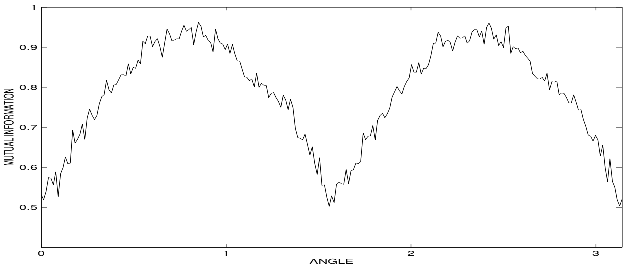

Fig. 5 Mutual information of $x'_1 \cos\alpha + x'_2 \sin\alpha$ and $-x'_1 \sin\alpha + x'_2 \cos\alpha$ as a function of α .

$$\hat{M} = \begin{bmatrix} -0.2084 & -0.8018 \\ 0.8052 & 0.2212 \end{bmatrix}$$

The quality of our ICA solution will now be compared to that of the algorithm FastICA, [7]. In particular, we test all six variants of the algorithm that differ in the minimization procedure (deflation/symmetric) and in the preselected nonlinear function (pow3/tanh/gauss). The results are listed in Table 1, where the first line refers to our procedure and the six other lines to FastICA. Two criteria are considered: 1) the mutual information of separated signals and 2) the error in estimating the mixing matrix. Note that

[1] The data were provided by courtesy of the Linguistic Data Consortium, http://www.ldc.upenn.edu/ .

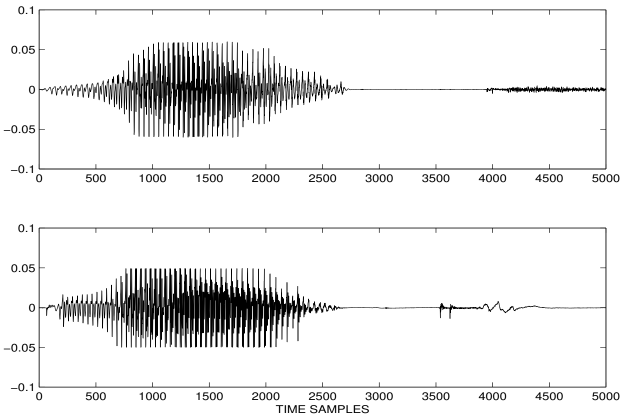

Fig. 6 Signals reconstructed from the unknown mixture.

TABLE 1: PERFORMANCE COMPARISON 1

	$I(\hat{s}_1, \hat{s}_2)$	$\sigma(\hat{M})$
mutual information	0.49	0.0222
deflation/pow3	0.57	0.0404
deflation/tanh	0.55	0.0734
deflation/gauss	0.62	0.0923
symmetric/pow3	0.58	0.0484
symmetric/tanh	0.55	0.0447
symmetric/gauss	0.54	0.0412

the mixing matrix can be estimated in general only up to order and signs of its rows. The error can be computed as

$$\sigma(\hat{M}) = \min_{\hat{M}' \equiv \hat{M}} \|\hat{M}' - M\| \tag{10}$$

where the minimization is performed over all matrices $\hat{M}'$ which differ from $\hat{M}$ only in the order and/or the sign of the rows. The matrix norm $\|\cdot\|$ is defined as the largest singular value.

Note that our algorithm based on direct minimization of the mutual information better separates the signals than FastICA, with respect to both criteria.

4. EXAMPLE 2: SEPARATION OF THREE SPEECH SIGNALS

The input data are three speech data files with the words "red", "green"and "blue".[1] The data file were both shortened so that both data sets have equal length, $N = 5000$. Let the data be arranged in an $(N \times 3)$ matrix denoted $s = [s_1, s_2, s_3]$. The signals are displayed in Figure 3. In this example, the mutual information of the signals was estimated by our algorithm to be $I(s) = 0.93$. The correlation matrix of the input signals is

$$\mathrm{corr}(s) = \begin{bmatrix} 1 & -0.004 & 0.0008 \\ -0.004 & 1 & 0.0988 \\ 0.0008 & 0.0988 & 1 \end{bmatrix}$$

We show that it is possible to recover the signals from an unknown linear mixture.

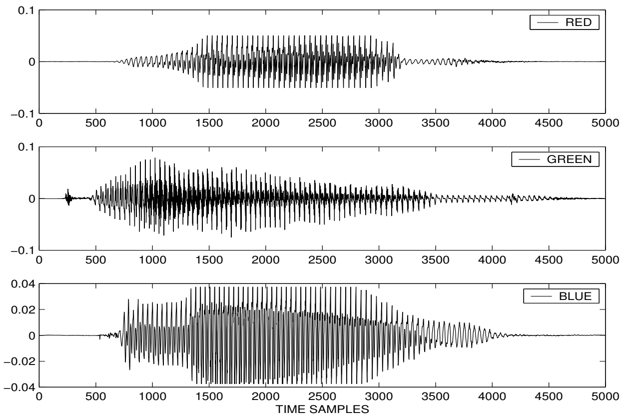

Fig. 7 Original speech signals.

Assume that the observed data are given as

$$x = s \cdot M \tag{11}$$

where M is a 3×3 matrix, to be estimated. In our example we used

$$M = \begin{bmatrix} 0.8 & 0.2 & 0.2 \\ 0.2 & 0.8 & 0.2 \\ 0.2 & 0.2 & 0.8 \end{bmatrix} . \tag{12}$$

The observed data (the mixtures) are plotted in Figure 8. The mutual information of the mixed signals is 1.83 .

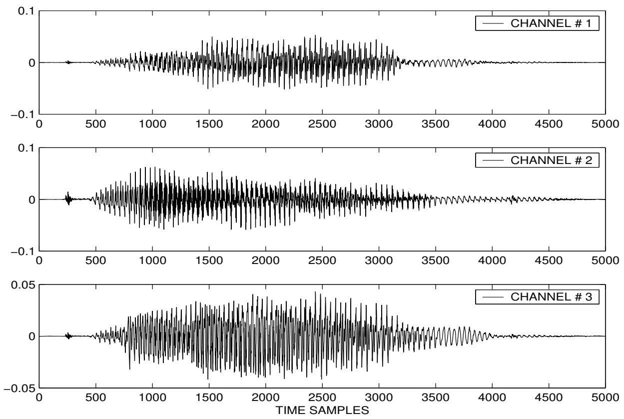

Fig. 8 Observed mixtures of the original signals.

As in the previous example, the signal separation could be performed by minimizing the mutual information of columns of the matrix xWQ, over the manifold $\mathcal{Q}$ of all orthogonal matrices of the dimension 3×3, where $W = (x^T x)^{-1/2}$ is the decorrelating matrix. This approach is, however, computationally very demanding. As in $\mathbb{R}^2$, the mutual information of xWQ as a function of Q is a nonsmooth function which has many local minima. Since any orthogonal transformation in the three-dimensional Euclidean space can be decomposed to three rotations, the minimization would proceed in $\mathbb{R}^3$.

Instead, we suggest an approximation, which appears to have a good performance and involves seven subsequent one-dimensional minimizations similar to that in the two-dimensional case. Let us introduce the following notation. For any $N \times 1$ vectors x_1, x_2 let

$$[y_1, y_2] = \mathrm{MMI}(x_1, x_2) \triangleq \arg\min_{\alpha} I(x(x^T x)^{-1/2} Q(\alpha))$$

denote two orthogonal vectors y_1, y_2 that minimize the mutual information of the two columns of $xWQ(\alpha)$. ("MMI" stands for Minimizing the Mutual Information.)

Our approximation is based on the hypothesis, that if x_1 and x_2 were given as two distinct linear combinations of three independent sources, s_1, s_2 and s_3, the optimally separated vectors y_1 and y_2 will be given as linear combinations of only two of the vectors. Say that y_1 will be a linear combination of s_1 and s_2, and y_2 will be a linear combination of s_2 and s_3.

Let us return to the original problem. We start with the prewhitened data, $x' = xW = x(x^T x)^{-1/2}$ and perform six one-dimensional searches:

$$\begin{aligned} [y_1, y_2] &= \mathrm{MMI}(x'_1, x'_2) \\ [z_1, z_2] &= \mathrm{MMI}(x'_1, x'_3) \\ [t_{1ij}, t_{2ij}] &= \mathrm{MMI}(y_i, z_j) \qquad i, j = 1, 2 \end{aligned}$$

It follows from our hypothesis that one of the pairs of vectors t_{1ij}, t_{2ij} which have the minimum mutual information represents two of the independent sources, $\hat{s}_1$ and $\hat{s}_2$. Without any loss of generality let us assume that the minimum is achieved for $i = 1$ and $j = 1$. The remaining third source can be identified as an orthogonal complement of t_{111}, t_{211} in the linear space spanned by x_1, x_2 and x_3. We can put

$$\begin{aligned} t'_{311} &= x_k - (t_{111}^T x_k) t_{111} - (t_{211}^T x_k) t_{211} \\ t_{311} &= t'_{311} / \| t'_{311} \| \end{aligned}$$

where k is selected from $\{1, 2, 3\}$. In practice, it appears that it is useful to control the mutual information between t_{311} and t_{111}, t_{211}. Hence, we suggest to accept as $\hat{s}_1$ one of the two vectors t_{111}, t_{211} that has the minimum mutual information with t_{311}. Without any loss of generality we can assume that it is t_{111}, and put

$$[\hat{s}_2, \hat{s}_3] = \mathrm{MMI}(t_{211}, t_{311}) \ .$$

As we can see, the whole procedure consists of 7 one dimensional searches and utilizes algorithm for computing the mutual information of two vectors. Indeed, it is interesting to check the mutual information of all three separated signals.

The mixing matrix M can be estimated as

$$\hat{M} = [\hat{s}_1, \hat{s}_2, \hat{s}_3]^T x$$

In our example we got

$$\hat{M} = \begin{bmatrix} -0.1731 & -0.2682 & -0.8070 \\ 0.8093 & 0.2069 & 0.2457 \\ 0.2048 & 0.7973 & 0.1998 \end{bmatrix}$$

which is quite good estimate of the true mixing matrix, up to a change of order of signs and raws of the matrix. The separated signals are plotted in Figure 9. The mutual information of the signals was 0.93, which equals the mutual information of the original (unmixed) signals.

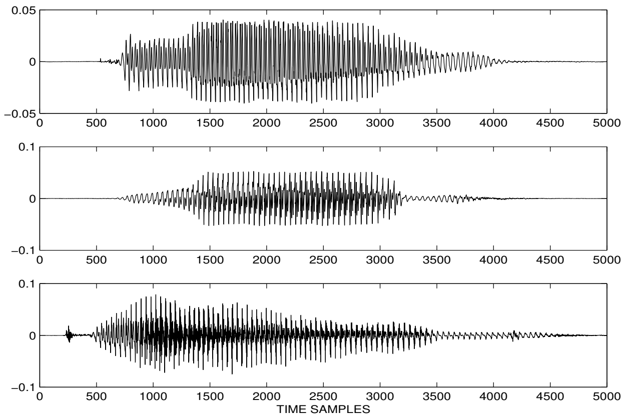

Fig. 9 Signals reconstructed from the unknown mixture.

TABLE 2: PERFORMANCE COMPARISON 2

	$I(\hat{s}_1, \hat{s}_2, \hat{s}_3)$	$\sigma(\hat{M})$
mutual information	0.93	0.0744
deflation/pow3	1.00	0.1007
deflation/tanh	1.01	0.0763
deflation/gauss	0.95	0.0638
symmetric/pow3	0.96	0.7588
symmetric/tanh	0.99	0.6433
symmetric/gauss	1.01	0.7198

A comparison of our procedure to the estimates provided by FastICA is given in Table 2. Note that the suggested algorithm better separates the signals than FastICA in terms of the achieved mutual information, and in five of the six cases it outperforms FastICA in terms of error of the mixing matrix estimate.

5. CONCLUDING REMARKS

The availability of good estimators of the mutual information, especially if they do not require any tuning, is important. The above algorithm for direct estimation of the mutual information allows to separate mixed signals without any special assumption about the distribution of the sources and without tuning. It also allow to check the quality of the separation.

6. REFERENCES

[1] C.R. Rao, "Principal Component and Factor Analyses", in G.S. Maddala and C.R. Rao eds., *Handbook of Statistics*, Vol. 14, , pp. 489-505, 1996.

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

[3] J.-F. Cardoso, "Blind signal separation: statistical principles", *Proceedings of the IEEE*, **9**(10), pp. 2009-2025, 1998.

[4] Te-Won Lee, *Independent Component Analysis, Theory and Applications*, Kluwer Academic Publishers, Boston, 1998.

[5] A. Hyvärinen, "Survey on independent component analysis", *Neural Computing Surveys*, Vol. 2, pp. 94-128, 1999.

[6] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis", *IEEE Trans. on Neural Networks*, Vol. 10, No.3, pp. 626–634, 1999.

[7] The FastICA MATLAB package. Available at http://www.cis.hut.fi/projects/ica/fastica.

[8] http://www.utia.cas.cz/user_data/scientific/SI_dept/Tichavsky.html .

[9] A. Hyvärinen, "New approximations of differential entropy for independent component analysis and projection pursuit". In *Advances in Neural Information Processing Systems 10* (NIPS*97), pp. 273-279, MIT Press, 1998.

[10] G.A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. Information Theory*, Vol. 45, no. 4, pp. 1315-1321, May 1999.

[11] G.A. Darbellay, "Predictability: an information-theoretic perspective." In: A. Prochazka, J. Uhlir, P.J.W. Rayner and N.G. Kingsbury (Eds), *Signal Analysis and Prediction*, pp. 249-262, Birkhauser, Boston, 1998.

[12] G.A. Darbellay, "An estimator for the mutual information based on a criterion for independence," *Journal of Computational Statistics and Data Analysis*, Vol. 32, pp.1-17, 1999.