Blind Signal Separation with a flexible non-linearity *

Hualou Liang and Nikos K. Logothetis
Email:{hualou.liang, nikos.logothetis}@tuebingen.mpg.de

Max-Planck Institute for Biological Cybernetics
Spemannstr. 38
72076 Tuebingen, Germany

ABSTRACT

Blind signal separation refers to the recovery of a set of independent sources when only mixtures of these sources with unknown coefficients are observed. This paper introduces a method for source separation without any knowledge of their probability distribution. This is achieved under a maximum likelihood framework by considering a simple parametric model which is constructed with a family of exponential power density function. As a result, an explicit algorithm for the adaptation of the non-linearity to various marginal densities in ICA is derived. Simulation result shows that our algorithm is able to separate a wider range of source signals, including mixtures of Gaussian sources. On real data, we demonstrate the successful use of our procedure to analyze ECG recordings.

KEYWORDS: Blind source separation, Independent Component Analysis, Latent variable density modeling

1. INTRODUCTION

In blind source separation or Independent Component Analysis(ICA) the problem is to recover sources that are considered statistically independent given the outputs of a number of sensors, each of which receives a different linear combination of the source signals. This kind of blind source separation technique is important in the signal processing area due to its prospective application in speech recognition, telecommunications and various medical signal processing applications.

Recently, blind source separation has become a highly popular research topic both in unsupervised neural learning and statistical signal processing [1, 2, 3, 4, 5, 6, 7, 8]. Researchers using neural networks for signal processing have considered the blind source separation problem from different approaches, including information theory, maximum likelihood estimation, higher order statistics, and nonlinear generalizations of Hebbian learning rules. For a review see [9, 10].

Despite its recent popularity, aspects of the ICA algorithms are still poorly understand. In this paper we seek to better understand independent component analysis through calculating the likelihood landscape. In specific, separation is accomplished by considering a maximum likelihood solution corresponding to some given distribution which is modeled by density functions of the exponential power family. As a result, an explicit algorithm for the adaptation of the nonlinearity to various marginal densities in ICA, which is solved by the numerical method, is derived. The results illustrate the performance of the method and the good agreement between simulations and theory.

2. THEORY

Let us consider unknown source signals $s_i(t), i = 1, \cdots, n$ which are mutually independent. It is assumed that the sources $s_i(t)$ are stationary processes. The mixture of the sources of the sensor output can be described by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{1}$$

where $\mathbf{A} \in \mathbf{R}^{n\times n}$ is an unknown non-singular mixing matrix, $\mathbf{s}(t) = [s_1(t), \cdots, s_n(t)]^T$ and $\mathbf{x}(t) = [x_(t), \cdots, x_n(t)]^T$.

Without knowing the source signals and the mixing matrix, we want to recover the original signals from the observations $\mathbf{x}(t)$ by the following linear transform:

$$\begin{aligned} \mathbf{u}(t) &= \mathbf{W}\mathbf{x}(t) \\ &= \mathbf{W}\mathbf{A}\mathbf{s}(t) \end{aligned} \tag{2}$$

where $\mathbf{u}(t) = [u_1(t), \cdots, u_n(t)]^T$ is an estimate of the sources and $\mathbf{W} \in \mathbf{R}^{n\times n}$ is a separating matrix. The sources are exactly recovered when $\mathbf{W}$ is the inverse of $\mathbf{A}$. After a possible permutation and scale change, we get the performance matrix $\mathbf{P}$

$$\mathbf{P} = \mathbf{W}\mathbf{A} \tag{3}$$

so that if $\mathbf{P}$ is normalized and reordered a perfect separation leads to the identity matrix.

Maximum Likelihood Estimation(MLE) approaches to ICA were first proposed in [11] and elaborated in [12]. The basic idea of ICA in term of maximum likelihood is to model the observation $\mathbf{x}$ as being generated from hidden variables $\mathbf{s}$ via a linear mapping $\mathbf{A}$. The likelihood of the data set in the given model is a function of the model's parameters. In the noiseless case, we can use a parametric density estimator $\hat{p}(\mathbf{x};\mathbf{a})$ to find the parameter vector $\mathbf{a}$ that minimizes the difference between the generative model $\hat{p}(\mathbf{x};\mathbf{a})$ and the observed distribution $p(\mathbf{x})$. Note that each vector $\mathbf{a}$ can be considered as a basis vector of the matrix $\mathbf{A}$ so that $\hat{p}(\mathbf{x};\mathbf{a})$ is an estimate of the observed vector $p(\mathbf{x})$. An appropriate difference between the estimate and the observation can be measured using the KL divergence, D:

$$\begin{aligned} D(p(\mathbf{x}), \hat{p}(\mathbf{x};\mathbf{a})) &= \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\hat{p}(\mathbf{x};\mathbf{a})} d\mathbf{x} \\ &= H(\mathbf{x}) - \int p(\mathbf{x}) \log \hat{p}(\mathbf{x};\mathbf{a}) d\mathbf{x} \end{aligned} \tag{4}$$

where $p(\mathbf{x})$ is the p.d.f. of the observation $\mathbf{x}$ and $\hat{p}(\mathbf{x};\mathbf{a})$ is a parametric estimate of the distribution $p(\mathbf{x})$. The divergence $D(p(\mathbf{x}), \hat{p}(\mathbf{x};\mathbf{a}))$ is zero only if our estimate $\hat{p}(\mathbf{x};\mathbf{a})$ matches the observation $p(\mathbf{x})$. The equivalence between the infomax

*This paper appears in the Proceedings of ICONIP 1998, Kitakyushu, Japan

principle [2] and maximum likelihood for independent component analysis is shown in [13]. The normalized log-likelihood of $\hat{p}(\mathbf{x};\mathbf{a})$ is therefore:

$$L = \frac{1}{T}\sum_{t=0}^{T}\log\hat{p}(\mathbf{x_t};\mathbf{a}) \tag{5}$$

where T is the number of observations of $\mathbf{x}$.

To obtain a maximum likelihood algorithm we find the gradient of the log likelihood through applying a stochastic gradient optimization. By introducing $\mathbf{W} = \mathbf{A}^{-1}$, the log-likelihood for a single observation $\mathbf{x}$ is given [12] by

$$\log p(\mathbf{x}|\mathbf{W}) = \log\det\mathbf{W} + \sum_i \log p_i(W_{ij}x_j) \tag{6}$$

We then obtain the following formula,

$$\frac{\partial D}{\partial \mathbf{W}} = -\mathbf{W}^{-T} - \frac{p_i'}{p_i}\mathbf{x}^{\mathbf{T}} \tag{7}$$

Setting the $p_i(u_i)$ to the derivative of the activation function of a logistic function, the algorithm reduces to that described in Bell and Sejnowski's paper [2]. Since $H(\mathbf{x})$ is not dependent on $\mathbf{W}$, maximizing the log likelihood is minimizing the KL divergence:

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{W}} &= -\frac{\partial D}{\partial \mathbf{W}} \\ &= \mathbf{W} + \frac{p_i'}{p_i}\mathbf{u}^{\mathbf{T}}\mathbf{W} \end{aligned} \tag{8}$$

In the above formula natural or relative gradient is used to enhance the computational efficiency.

When calculating the likelihood, an open question is how to select the form of the activation function so that it can match input's pdf. In other words, this is how to choose the marginal density functions, $p_i(u_i)$, which in practice, however, are not known. Here we consider the family of exponential power density functions to be our set of basis function:

$$p_\alpha(x) = c_1\exp(c_2|x|^\alpha) \tag{9}$$

where α is a positive constant, and c_1, c_2 are normalization constants that ensure that $p_\alpha(x)$ is a probability density of unit variance. The width of the distribution is controlled by c_2, and the weight of its tails is determined by α. When α is set to different values, we obtain different densities which range from super-Gaussian(positive kurtosis) to sub-Gaussian(negative kurtosis). For example, p is Gaussian when $\alpha = 2$. Thus the densities in this family can be used as examples of different non-Gaussian densities. This is why we choose this set of the non-linearity function.

An alternative to the parametric density model above, which may also be used for the separation of sub- and super-Gaussian sources, can be defined by

$$p(x) = \mu\varphi(x) + (1-\mu)c\varphi(c(x-1))) \tag{10}$$

where c is constant, and μ is a parameter that takes all the values in the interval $0 \le \mu \le 1$, $\varphi(x)$ is the Gaussian function. In fact, this is the case of mixtures of Gaussian, which was discussed in [14]. This family includes asymmetric densities of both negative and positive kurtosis. Only the exponential power family of density functions will be considered in this context. Similar derivation should be obtained for Gaussian mixture densities.

The normalised log likelihood is

$$L = \frac{1}{T}\sum_{t=1}^{T}\log p_\alpha(x) = \log c_1 + c_2|x|^\alpha \tag{11}$$

To learn the parameter α, for example, we need the derivative

$$\frac{dL}{d\alpha} = c_2|x|^\alpha\log|x| \tag{12}$$

To compute the parameters c_1, c_2 and α of L (see equation(11) and update $\mathbf{W}$(see equation(8)), schemes based on the natural gradient algorithm [4] and the quasi-Newton method which have been implemented can be used. The simulation given below is based on the quasi-Newton method, in which the precision required of the objective function of minimizing $-\log L$ is set to 10^{-6}, the maximum number of iterations is 100. Here the precision is defined as the absolute difference between the objective function values of two successive steps. One complication is to note that, after each weight update, the weights $\mathbf{W}$ must be normalized, $\sum_j W_{ij}^2 = 1$.

3. SIMULATION

In order to check the validity and performance of our approach, both simulation experiment and read world example are given below. First, we simulate it on computer using synthetic source signals and random mixing matrix. Second, we demonstrate it on separating the fetal electrocardiogram(ECG) from the maternal trace and from the other contaminations.

3.1. Simulation Experiment

Six source signals that having visually appealing waveform are randomly mixed. It should be noted, unlike the source signals used in other experiments which usually consist of only *one* Gaussian, we use two Gaussian. The first 100 values of the source signals, whose total length was 2000 points, are depicted in Fig. 1. From top to bottom of the left column, there are two sub-Gaussian, two super-Gaussian and two Gaussian respectively. The mixed signals are plotted in the middle column, and the separation result is shown in the right column.

But one question still leaves open - what should the criteria be for two Gaussian signals to be considered as separated. To address this question we calculate the correlation coefficients between source signals and separated ones, which are given by the following matrix,

$$\begin{bmatrix} -0.004 & 0.038 & -0.002 & 0.999 & 0.004 & 0.001 \\ 0.069 & -0.028 & \mathbf{-0.997} & 0.0004 & 0.001 & 0.0002 \\ \mathbf{0.999} & -0.007 & 0.046 & 0.003 & -0.011 & 0.003 \\ -0.030 & \mathbf{-0.997} & 0.036 & 0.052 & -0.037 & 0.010 \\ -0.046 & 0.064 & 0.008 & 0.010 & \mathbf{-0.997} & 0.007 \\ 0.007 & -0.009 & -0.003 & 0.032 & -0.016 & \mathbf{-0.999} \end{bmatrix}$$

It can be seen from the above source correlation matrix that the well separation is achieved by our method.

The performance of the algorithm can also be examined in another way. Fig. 2 shows the performance matrix $\mathbf{P}$ as indicated in equation(3). If the separation is perfect, the result will be a ridge along the diagonal with all off-diagonal elements close to zero.

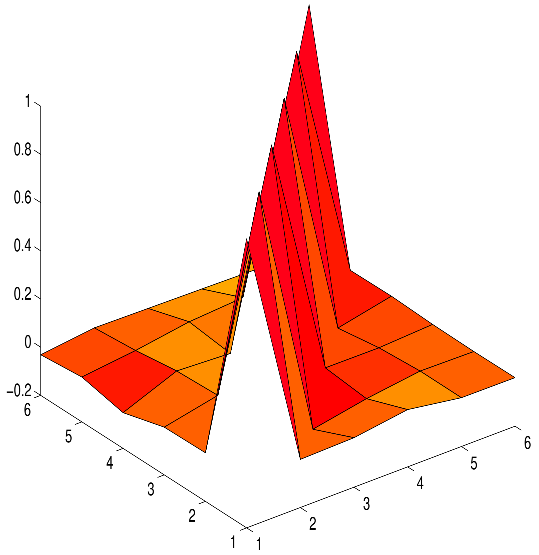

Figure 2: Performance matrix for the separation of 6 sources.

3.2. Experimental Result on ECG

The ECG data under analysis are 8-channel cutaneous potential recordings of a 16 weeks old fetus. Electrodes were put on the abdomen and on the cervix of the mother. Fig. 3 (a) shows 10 seconds segment of the recordings resampled at 500 Hz. From the result of blind signal separation displayed in Fig. 3 (b), fetal and maternal trace can be clearly distinguished. Also the noise contaminations and breathing signal are well separated.

4. DISCUSSION AND CONCLUSIONS

Ours is not the first attempt at modeling a parametric nonlinearity. The idea of using a parametric model to match the source distribution was already recognised in [2], which was further generalized in the contextual ICA(cICA) algorithm [12] by Pearlmutter and Parra. They model the p.d.f. in a parametric form by taking into account the temporal information and by considering it as a weighted sum of several logistic density functions with variable means and scales. Xu et al. [14] model the underlying p.d.f. with mixtures of Gaussian and show that they can separate sub and super-Gaussian sources. And more recently, Lee et al. [9] use single parametric model and a stability criterion to achieve separation of sub and super-Gaussian sources.

In view of this previous work, what do we think is the contribution of the present article ?

- From the theoretical point of view, the problem of blind signal separation is formulated within the framework of maximum likelihood, in which the marginal densities are explicitly modeled by the exponential power family of density functions.
- From the operational point of view, we explicitly calculate the likelihood function during learning the nonlinearities.
- Further, this technique is able to capture some salient features of the underlying signal distribution, enabling separation of mixture of Gaussian sources which usually may not be separated by other algorithms.

Numerical simulation shows our approach allows to separate a wider variety of sources ranging from sub-Gaussian to super-Gaussian, and even the mixture of Gaussian provided only that they have distinct spectra. The success of our procedure is further demonstrated on real world application of ECG signal separation.

References

[1] Comon, P., "Independent Component Analysis, a new concept ?," Signal Processing, 36:287-314, (1994).

[2] Bell, A. and Sejnowski, T., "An information-maximisation approach to blind separation and blind deconvolution," Neural Computation, 6:1129-59 (1995).

[3] Cardoso J-F. and Laheld B., "Equivariant adaptive source separation," IEEE Trans. on Signal Proc., 45, 2, 434-444 (1996).

[4] Amari, S., Cichocki, A., and Yang, H., "A New Learning Algorithm for Blind Signal Separation," In Advances in Neural Information Processing Systems 8, pp.757-763 (1996).

[5] Amari, S., "Natural gradient works efficiently in learning." Neural Computation, in press.

[6] Girolami, M., Self-organizing artificial neural networks for signal separation, Ph.D thesis, Department of Computing and Information Systems, Paisley University, Scotland, (1997).

[7] Oja E., "The nonlinear PCA learning rule in Independent Component Analysis," Neural Computation, 17, 25-45 (1997).

[8] Lee T-W., Girolami M., Bell AJ and Sejnowski TJ, "Independent Component Analysis using an Extended Infomax Algorithm for Mixed Sub-Gaussian and Super-Gaussian sources," Neural Computation, (1998).

[9] Lee T-W., Girolami M., Bell AJ and Sejnowski TJ, "A Unifying Information-theoretic Framework for Independent Component Analysis," International Journal on Mathematical and Computer Modeling, in press.

[10] Cardoso J-F., "Blind signal separation: a review," Proc. of IEEE, in press.

[11] Gaeta M. and Lacoume J., "Source separation without a priori knowledge: The maximum likelihood solution," Eusipco 90, Barcelone, 621-624 (1990).

[12] Pearlmutter B.A. and Parra L.C. 1996. "A context-sensitive generalization of ICA," Proc. ICONIP '96, 151-157 (1996).

[13] Cardoso J-F., "Infomax and maximum likelihood for blind separation," IEEE Signal Processing Letters, 4, 109-111 (1997).

[14] Xu, L., Cheung, C., Yang, H. and Amari, S., "Maximum equalization by entropy maximization and mixture of cumulative distribution functions." IEEE Proc. ICNN, Houston, 1821-1826 (1997).

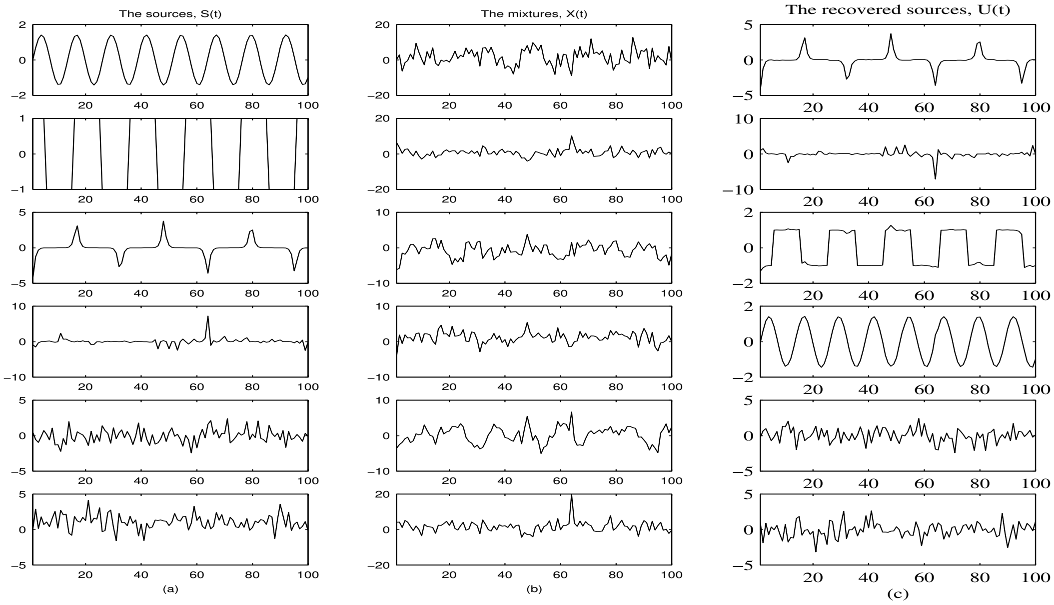

Figure 1: Example of separation of independent source signals. (a) The sources. (b) The set of random mixtures of source signals in (a). (c) Estimates of the source signals.

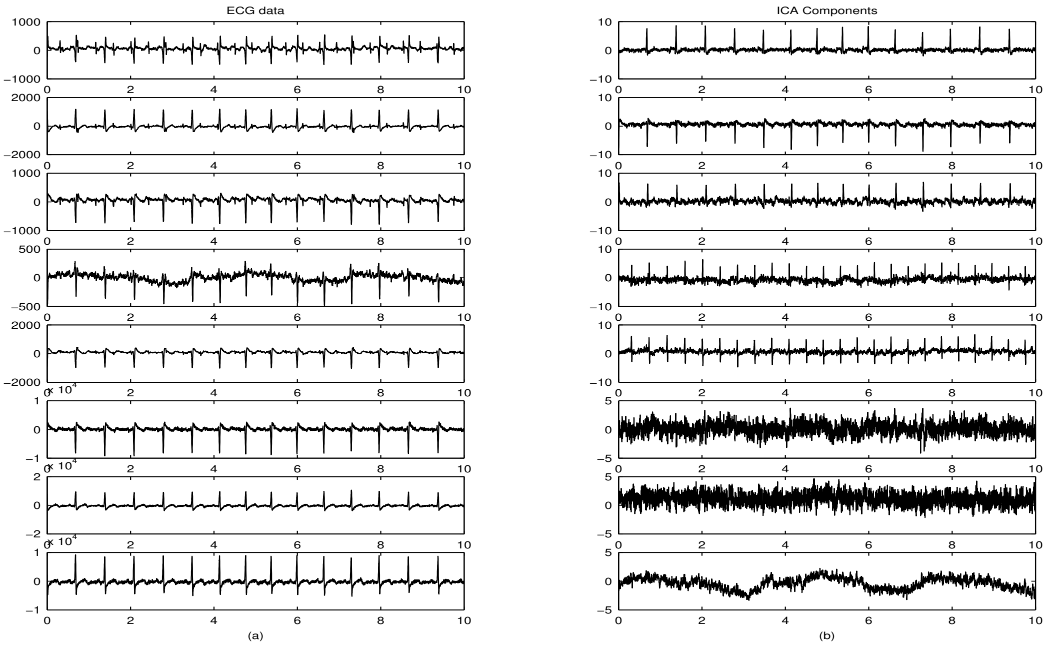

Figure 3: Example of separation of independent source signals of ECG data. (a) Noninvasive ECG recordings. (b) The 8 components extracted from ECG data. The first 3 components and its next 2 components are mother and fetal traces respectively. The last trace is the breathing signal and others are noise contaminations.