Adaptive Independent Component Analysis through a Maximum Likelihood Approach *

Hualou Liang and Nikos K. Logothetis
Max-Planck Institute for Biological Cybernetics
Spemannstr. 38, 72076 Tuebingen, Germany
E-mail: {hualou.liang, nikos.logothetis}@tuebingen.mpg.de

Abstract

In this paper we propose a method for separating mixtures of independent sources without any precise 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. We demonstrate that our algorithm is able to separate a wider range of source signals, including mixtures of Gaussian sources.

1 Introduction

Blind source separation or Independent Component Analysis(ICA) is a rapidly emerging, promising technique in the field of neural networks and statistical signal processing. For a review see [1]. Despite its popularity in recent theoretical and empirical development [2, 3, 4, 5, 6, 7, 8], aspects of the ICA algorithms are still largely unexplored. 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

One popular way for formulating the ICA problem is to consider the parametric estimation of the following generative model for the data:

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

where $\mathbf{s}(t) = [s_1(t), \cdots, s_n(t)]^T$ is an n-dimensional random vector whose components are assumed mutually independent, $\mathbf{x}(t) = [x_(t), \cdots, x_n(t)]^T$ is an observed n-dimensional vector, and $\mathbf{A} \in \mathbf{R}^{n\times n}$ is an unknown non-singular mixing matrix.

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:

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

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.

The basic idea of ICA in term of maximum likelihood is to model the observation $\mathbf{x}$ as being generated from latent 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

*This paper appears in the Proceedings of ICNN & B 1998, Beijing, China

if our estimate $\hat{p}(\mathbf{x};\mathbf{a})$ matches the observation $p(\mathbf{x})$. The equivalence between the infomax principle [3] and maximum likelihood for independent component analysis is shown in [9]. 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 [10] 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,

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

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 [3]. Since $H(\mathbf{x})$ is not dependent on $\mathbf{W}$, maximizing the log likelihood is minimizing the KL divergence:

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

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.

The normalised log likelihood is

$$L = \frac{1}{T}\sum_{t=1}^{T} \log p_\alpha(x) = \log c_1 + c_2|x|^\alpha \tag{10}$$

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

$$\frac{dL}{d\alpha} = c_2|x|^\alpha \log|x| \tag{11}$$

To compute the parameters c_1, c_2 and α of L (see equation(10) and update $\mathbf{W}$(see equation(8)), schemes based on the natural gradient algorithm [5] 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 Results

In order to check the validity and performance of our approach, the extensive computer simulations are given below. First, we simulate it on computer using synthetic source signals and random mixing matrix. Second, we demonstrate it on restoring images after degradation and interfering images.

3.1 Simulation 1

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 **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.

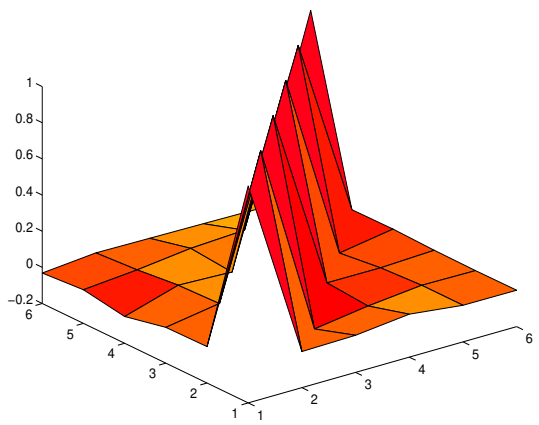

Figure 2: Performance matrix for the separation of 6 sources.

3.2 Simulation 2

Seven natural images and one uniformly distributed noise were chosen as the sources. All images have the same size of 256×256 pixels, each pixel using an eight-bit greyscale representation. The image sources and their probability density distributions are shown in the first two rows in Figure 3. The mixed images and their histogram distributions shown in the middle two rows in Figure 3 were generated by a randomly mixing matrix of 8×8 in size. Table 1 gives the source and mixed image statistics. It is clear to see from Table 1 and Figure 3 that the selected source images have both sub- and super-Gaussian distributions, and the mixing has caused significant degradation of the images, making them indistinguishable. The source images which are shown in the last two rows in Figure 3 are well separated, even there is still a little degradation to some extent.

Table 1: Original and mixed images statistics

Image	Kurtosis	Mixed Images	Kurtosis
Einstein	0.4394	Mix1	-0.0577
Reagan	-1.2463	Mix2	-0.0983
Lenna	-0.7584	Mix3	-0.1718
Hat	-0.9490	Mix4	-0.5124
Girl	-0.9549	Mix5	-0.0459
Oleh	-1.4347	Mix6	-0.2103
Kids	2.8787	Mix7	-0.4123
Noise	-1.2058	Mix8	-0.2991

4 Concluding remarks

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.

Numerical simulations show our approach allows to separate a wider variety of sources ranging from sub-Gaussian to super-Gaussian, and even the mixture of Gaussian which usually may not be separated by other algorithms.

References

[1] 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.

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

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

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

[5] 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).

[6] 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).

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

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

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

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

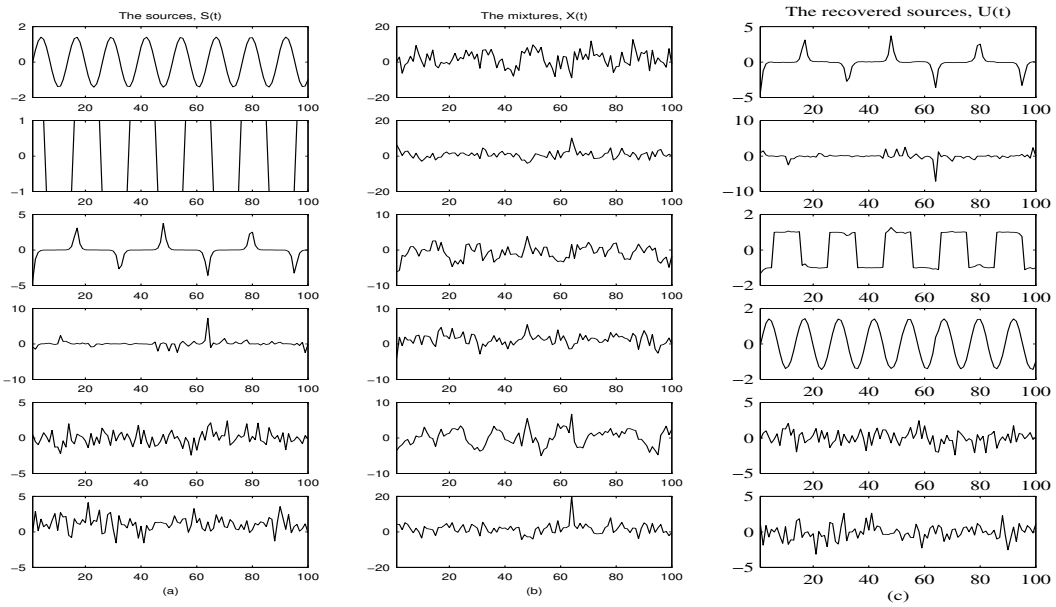

Figure 1: Example of separation of independent source signals. (a) The sources. (b) The set of random mixtures of source signals in (a). (c) Estimates of the source signals.

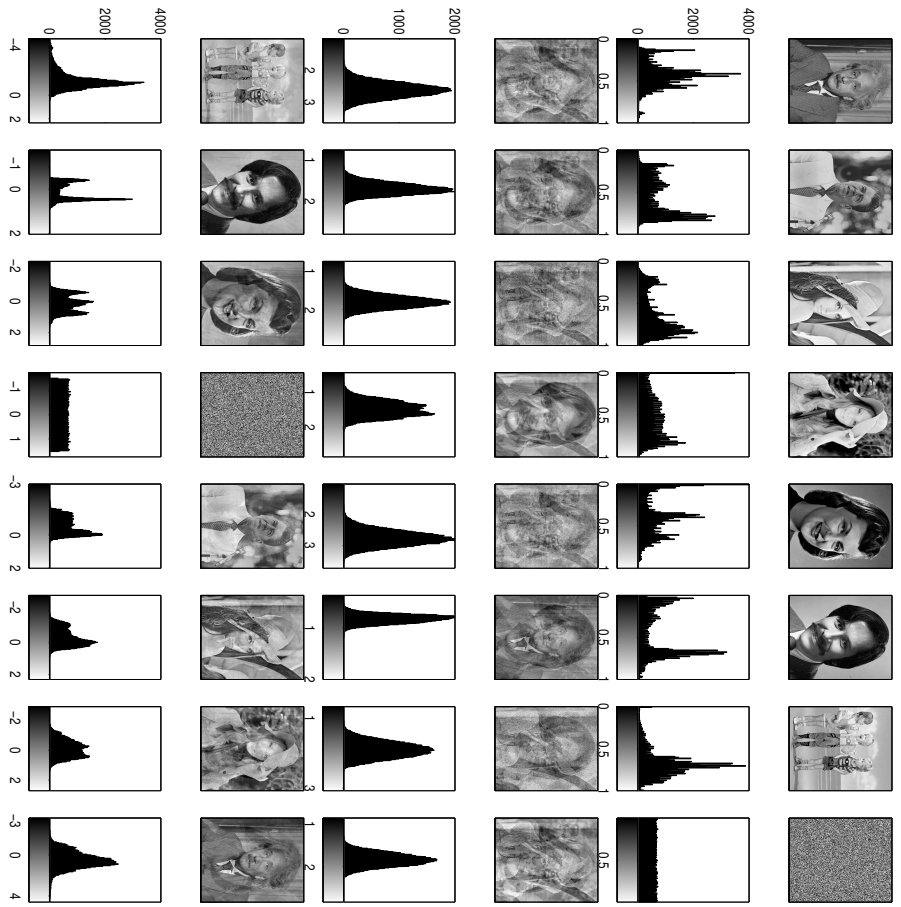

Figure 3: Separation of mixed images. The first two rows show the original source images and their histogram distributions; those in the middle two rows are for the mixed images; and the last two rows give the separated images and their histogram distributions.