A NEW ICA ALGORITHM FOR BLIND SOURCE SEPARATION

Chin-Jen Ku, Terrence L. Fine

Department of Electrical and Computer Engineering
Cornell University. Ithaca, New York 14850. U.S.A.
e-mails: kucj@ece.cornell.edu, tlfine@ece.cornell.edu

ABSTRACT

In this paper we propose a nonparametric independent component analysis (ICA) algorithm for the case of instantaneous and linear mixtures. Our algorithm combines minimization of correlation among nonlinear expansions of the output signals with good initialization derived from search guided by statistical tests for independence. Simulation results obtained from both synthetic and real-life data show that our method provides consistent result and compares favorably to the existing ICA algorithms.

1. INTRODUCTION

The problem of blind source separation (BSS) arises when one desires to extract certain signals based solely on their mixture(s). We present a new nonparametric BSS algorithm enabling reliable recovery of the original signals by iteratively minimizing the correlation among nonlinear expansions (NLExp) of the output signals. A key to our success is finding a suitable initial starting point by implementing an efficient statistical independence test. Our study is focused on the problem of source separation for *instantaneous* and *time-invariant linear* mixtures. Let us first define N as the number of sources, M the number of sensors (i.e. received signals) and L the number of samples. One may express the mixing process as

$$X = A \cdot S \tag{1}$$

where S is an $N \times L$ matrix representing samples of N independent sources, X is an $M \times L$ matrix of observed mixtures and A is an $M \times N$ real constant matrix with rank N ($M \geq N$), usually called the *mixing matrix*. Without loss of generality, we assume that the sources are zero-mean. The demixing process recovers the N independent source signals in S as Y by multiplying a demixing matrix W by the observed mixtures X:

$$Y = WX = WAS = CS \tag{2}$$

The source separation is successful if the matrix C rescales and/or permutes the rows of S. We assume absence of noise and present the case of equal number of sources and received signals (N=M) that is known at the processing unit. BSS problem can be translated into *Independent Component Analysis* (ICA) for the ultimate goal consists of recovering statistically independent outputs.

2. NLEXP-ICA ALGORITHM

It is well known from probability theory that two scalar random variables X_1, X_2 are statistically independent if and only if $f(X_1)$ and $g(X_2)$ remain uncorrelated for all f and g ranging over a separating class of functions [7]. Hence, we may define a measure of dependence based on the residual correlation among $f(X_1)$ and $g(X_2)$ with adequate selection of the functions f, g. Given a demixing matrix W and the corresponding output signals Y, our new algorithm first expands Y to $Z_1, Z_2, \ldots Z_F$ where $Z_i = f_i(Y)$ such that $f_1(Y) = Y$ and f_i being some nonlinear functions applied elementwise to Y for $i \geq 2$. This results in the following $NF \times L$ compound matrix

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_F \end{pmatrix} = \begin{pmatrix} f_1(Y) \\ \vdots \\ f_F(Y) \end{pmatrix} \tag{3}$$

Its covariance matrix is given in block form by

$$C_{\mathbf{Z}} = \begin{pmatrix} Cov(Z_1, Z_1) & \ldots & Cov(Z_1, Z_F) \\ \vdots & & \vdots \\ Cov(Z_F, Z_1) & \ldots & Cov(Z_F, Z_F) \end{pmatrix} \tag{4}$$

In order to take into account the cross-covariances only with respect to pairs of distinct variables $\{(y_i, y_j) : i \neq j\}$, we first form the matrix $\mathcal{M}$ defined as

$$\mathcal{M} = M \otimes C_{\mathbf{Z}} \tag{5}$$

The operator $\otimes$ denotes elementwise matrix multiplication and $M = \mathbf{1} - \mathbf{I}$, where $\mathbf{1}$ is an $NF \times NF$ matrix with all entries equal to one and $\mathbf{I}$ is an $NF \times NF$ matrix comprised of F^2 $N \times N$ identity matrices:

$$M = \mathbf{1} - \begin{pmatrix} I_N & \dots & I_N \\ \vdots & & \vdots \\ I_N & \dots & I_N \end{pmatrix} \tag{6}$$

We select as our objective function the sum of squared covariances of all pairs of variables $(f_p(y_i), f_q(y_j))$ for $i \neq j$ and express it as

$$\mathcal{L} = \frac{1}{2} \cdot trace(\mathcal{M} \cdot \mathcal{M}), \tag{7}$$

It can be shown that the gradient of $\mathcal{L}$ with respect to W is given by

$$\nabla_W \mathcal{L} = \frac{1}{L} \sum_{i,j=1}^{NF} \mathcal{M}_{ij} \cdot \mathcal{K}_{ij} \cdot X^T, \tag{8}$$

where

$$\mathcal{K}_{ij} = \underline{\mathbf{e}}_p DZ(i,\cdot) \otimes Z(j,\cdot) + \underline{\mathbf{e}}_q Z(i,\cdot) \otimes DZ(j,\cdot) \tag{9}$$

$$DZ = \begin{pmatrix} D_1 \\ \vdots \\ D_F \end{pmatrix} = \begin{pmatrix} f_1'(Y) \\ \vdots \\ f_F'(Y) \end{pmatrix} \tag{10}$$

and $\underline{\mathbf{e}}_i$ is the i^{th} column of identity matrix I_N. Thus the demixing matrix W may be estimated iteratively using standard gradient-based learning rules (*e.g.* steepest descent):

$$\hat{W}(k+1) = \hat{W}(k) - \eta(k) \cdot \nabla_W \mathcal{L}(k) \tag{11}$$

3. TESTS FOR STATISTICAL INDEPENDENCE

3.1. Power-divergence family test statistics

In discrete multivariate analysis, it is known that a hypothesis regarding the statistical properties among N discrete variables $y_1 \dots y_N$ can be tested by evaluating the chi-square goodness-of-fit statistics on the corresponding contingency table. Assume that the i^{th} variable y_i has been quantized into q intervals or cells $\{J_j^{(i)} : j = 1 \dots q\}$ for $i = 1 \dots N$. We define the *cell counting variables*

$$N_{\mathbf{j}} = N(\{j_1, j_2 \dots j_N\}) \tag{12}$$

as the number of data points where we have concurrently $y_1 \in J_{j_1}^{(1)}, \dots, y_N \in J_{j_N}^{(N)}$, with the argument $\mathbf{j}$ taking values in $\mathcal{J} = \{1 \dots q\}^N$. Read and Cressie [8] introduced the *power-divergence (PD) family* of goodness-of-fit statistics to measure the deviation of observed data from the stipulated hypothesis:

$$2I^{\lambda}(\mathbf{X}:\mathbf{m}) = \frac{2}{\lambda(\lambda+1)} \sum_{\mathbf{j}\in\mathcal{J}} N_{\mathbf{j}} \cdot [(\frac{N_{\mathbf{j}}}{m_{\mathbf{j}}})^{\lambda} - 1] \tag{13}$$

where $m_{\mathbf{j}}$ is the expected number of data belonging to the cell $\mathbf{j}$ under the hypothesis and $N_{\mathbf{j}}$ is the number of data actually observed. When the expected cell counts $\{m_{\mathbf{j}}\}$ are unknown they are replaced by their estimates $\{\hat{m}_{\mathbf{j}}\}$ based on the observed data. Under the assumptions of independence among the data points and fixed cell probabilities for all L, all the PD family statistics are asymptotically equivalent when the hypothesis is true. The common asymptotic distribution is χ^2 with a number of degrees of freedom depending on the total number of cells and the number of estimated parameters. Numerous researchers suggest that choosing λ between 0 and 1 (*e.g.* $\lambda = \frac{2}{3}$) has the best detection power against arbitrary lack of fit to the hypothesis.

3.2. Adaptation to ICA problems

The test statistics described above may be used to determine whether a given demixing matrix recovers statistically independent signals even when the individual signal samples $\{s_i(t)\}$ are not *i.i.d.* If the signals $\{s_i(t)\}$ are stationary and mutually independent of each other, the marginal sums of cell counting variables $\{N_{i+}, N_{+j}\}$ determine the unbiased estimates of expected cell counts $\mathbf{m}$. For $N = 2$, we have

$$\hat{m}_{ij} = \frac{N_{i+} \cdot N_{+j}}{L}, \quad (i,j) \in \{1, 2 \dots q\}^2 \tag{14}$$

with $N_{i+} = \sum_{j=1}^{q} N_{ij}$ and $N_{+j} = \sum_{i=1}^{q} N_{ij}$. When the individual signal samples are i.i.d, $\{N(\mathbf{j})\}$ are distributed according to a multinomial distribution and the estimates in (14) become the maximum likelihood estimates of $\mathbf{m}$.

Extension of equation (14) to higher dimensional cases is straightforward. Moreover, it is advised by many to define *equiprobable* cells to guarantee the unbiasness of PD test statistics [8]. In this case, the estimates of expected cell counts are equal to the ratio between sample size and the total number of cells:

$$\hat{m}_{\mathbf{j}} = \frac{L}{q^N} \quad \forall \mathbf{j} \in \mathcal{J} \tag{15}$$

The independence test can be easily performed by substituting the value of $\hat{\mathbf{m}}$ from (15) into (13). The exact number of degrees of freedom of the approximating chi-square distribution is equal to $(q^N - qN + N - 1)$ by taking into account all the constraints on the marginal sums.

3.3. Improving robustness of NLExp-ICA

Preliminary simulation results show that our *NLExp-ICA* algorithm sometimes encounters convergence difficulty or spurious solutions when the initial guess is far away from the true demixing solution. To improve its robustness, we propose a two-phase ICA algorithm. The first phase uses a power divergence statistics-based independence test to obtain a suitable initialization of demixing matrix W_0, whereupon *NLExp-ICA* refines the solution to reach a demixing matrix with high separation performance. Search for W_0 may be carried out using an orthogonal ICA algorithm [3] which is based on the eigen-decomposition of the covariance matrix of the observed mixtures:

$$R_{\mathbf{x}} = E\{\mathbf{x}\mathbf{x}^{\mathbf{T}}\} = AR_{\mathbf{s}}A^T = ELE^T \tag{16}$$

where $R_{\mathbf{s}}$ is diagonal under hypothesis of statistical independence among the sources, E is an orthogonal matrix containing the eigen-vectors of R_X, and L is a diagonal matrix with all positive entries on the diagonal since A is assumed invertible. Comparing the last two terms we conclude that the mixing matrix can be written as $EL^{\frac{1}{2}}UR_{\mathbf{s}}^{-\frac{1}{2}}$ where U is any orthogonal matrix. In virtue of the scaling ambiguity in blind identification [3], one can estimate the demixing matrix W by $U^T L^{-\frac{1}{2}} E^T$ so that $Y = WX$ recovers S to within a rescaling of rows by $R_{\mathbf{s}}^{-\frac{1}{2}}$ and a permutation and sign change determined by the choice of U. Hence, the problem is reduced to finding an orthogonal transformation U such that the resulting output Y yields low PD test statistics. The corresponding demixing matrix is used to initialize the *NLExp-ICA* algorithm.

4. EXPERIMENTAL RESULTS

4.1. Performance evaluation of ICA algorithms

We introduce a new index measuring the quality of source separation based on the *signal-to-interference energy ratio* (SIR) defined as follows. Let C be the global transfer matrix between original and estimated sources as defined in (2), Es_i^2 be the energy of i^{th} original source and $E_{i,j}$ the energy of i^{th} source recovered on j^{th} channel. The quantity

$$\gamma_{j,i} = \frac{E_{i,j}}{\sum_k E_{k,j}} = \frac{C_{ji}^2 Es_i^2}{\sum_k C_{jk}^2 Es_k^2} \tag{17}$$

represents the ratio between the energy of i^{th} source recovered on channel j and the total energy recovered on channel j. Now let

$$SIR_i(dB) = 10\log_{10}\frac{\gamma_i}{1-\gamma_i},\ \ \gamma_i = \max_j \gamma_{j,i} \tag{18}$$

then SIR_i is the ultimate quality of recovery with respect to the i^{th} source. If $SIR_i > 0$, the i^{th} signal is dominant at some output channel of the demixer. Conversely, if $SIR_i < 0$, the i^{th} signal does not dominate at any of the output channels. Such case happens when the recovery is *incomplete*. We will conservatively use the minimum SIR over all the channels (MSIR) as the global separation performance index. This definition is consistent with another standard performance index called *Amari error* [4] since MSIR increases when the matrix C tends toward an ideal demixing matrix (*i.e.* product of scaling and permutation matrices). However, unlike Amari error, the index MSIR is capable of detecting incomplete recovery.

4.2. Simulation setup

We illustrate the performance of our new algorithm on four different data sets. Data set 1 is generated synthetically, consisting of one *i.i.d.* Rayleigh(1) distributed random sequence and one *i.i.d.* Laplacian(1) distributed sequence. The three other data sets are taken from ICALAB benchmark sets [6]: *Speech4* (sources 1,2), *Sergio7* (sources 3,6) and *Gnband* (sources 1,4). Given the large size of data set 3 (L=10^4), we select the first 2000 samples from each source as the original signals. NLExp-ICA uses the original data plus the three following nonlinear expansions to form the compound data Z:

$$f_2(x) = \frac{1}{1+e^{-x}}, f_3(x) = \cos 3x, f_4(x) = \sin 2x \tag{19}$$

The performance of NLExp-ICA is compared to three other standard ICA algorithms: InfoMax, Fast-ICA (pow3) and JADE [1, 2, 5]. All the experiments conducted on the data sets are repeated over 200 randomly selected mixing matrices A and the average MSIR (dB) are reported.

4.3. Separation Performance

In Figure 1 we first illustrate the behavior of the PD test statistics in terms of the quality of demixing matrix W (measured by MSIR) for a simple two sources BSS problem. The orthogonal matrix U has one parameter (θ). We see readily that as the degree of separation improves, the PD test statistic decreases and stays below significance level when the separation is maximal. Figure 2 shows the performance realized by each of the four ICA algorithms on data set 1. Our algorithm clearly demonstrates better performance. In addition, we noted that InfoMax and Fast-ICA algorithms yielded unsatisfactory results for several randomly chosen mixing matrices as the resulting MSIR was below 5dB.

The results for data sets 2,3,4 are given in Table 1. We note in particular the performance realized by NLExp-ICA on data set 4, which contains two fourth order colored sources with a distribution close to Gaussian and is known to be a "hard" BSS benchmark. NLExp-ICA achieves much better separation than Fast-ICA and JADE. Overall, our algorithm succeeds in finding consistent demixing solution with superior performance compared to the three other ICA algorithms.

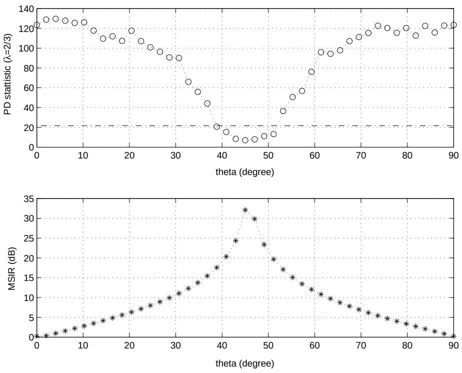

Fig. 1. Behavior of PD test statistic vs. quality of demixing matrix W. PD statistic is below significance level at maximal degree of separation (highest MSIR).

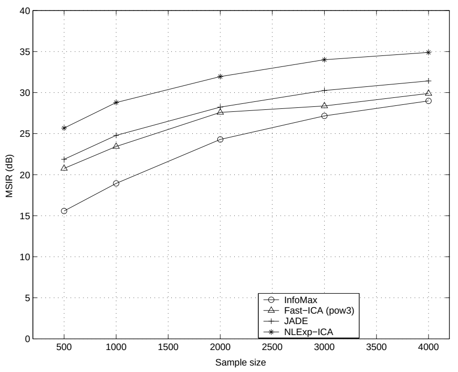

Fig. 2. BSS performance results for data set 1. Plot of average MSIR versus sample size L over 200 randomly selected mixing matrices A.

5. CONCLUSION

We have proposed a powerful nonparametric two-phase ICA algorithm. It first deploys a statistical independence test to find an initial estimate of demixing matrix achieving sufficient degree of independence among the recovered signals. Subsequently, it refines the solution by minimizing the correlation among different nonlinear expansions of all recovered signals. Experimental results show that our method provides consistent results for a variety of source distributions and can outperform the existing state-of-the-art ICA algorithms in some difficult problems. Furthermore, the proposed independence test may be regarded as a fast and efficient tool to evaluate the quality of solution provided by any ICA algorithm without knowing the sources. Our future investigation will be focused on the generalization of NLExp-ICA algorithm to high dimensional problems where time-efficient search of good initialization is needed.

Table 1. BSS performance results for data sets 2,3,4. Average MSIR (dB) versus sample size L.

	L	InfoMax	FastICA	JADE	NLExp
Data 2	500	20.66	31.26	31.37	**34.85**
	1000	30.29	32.53	35.08	**36.52**
	2000	19.44	24.09	21.19	**28.49**
	3000	29.38	32.44	26.23	**38.81**
	4000	31.46	**34.01**	30.58	33.38
	5000	30.97	31.92	31.58	**33.47**
Data 3	250	19.47	20.95	18.98	**30.08**
	500	28.11	36.49	39.04	**44.83**
	1000	33.26	36.28	40.36	**50.15**
	1500	34.54	29.71	31.65	**39.22**
	2000	35.05	31.88	35.22	**38.51**
Data 4	2000	13.12	5.32	5.43	**18.56**

6. REFERENCES

[1] Jean-Francois Cardoso, “Blind separation of instantaneous mixtures of nonstationary sources,” *IEEE Transactions on Signal Processing*, vol. 49, no. 9, September 2001.

[2] Aapo Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626-634, 1999.

[3] Jean-Francois Cardoso, “Blind signal separation: Statistical Principles,” *Proceedings of the IEEE*, vol. 86, no. 10, October 1998.

[4] S. Amari, A. Cichocki and H. H. Yang, “A new learning algorithm for blind signal separation,” *Advances in Neural Information Processing Systems 8*, MIT Press, 1996.

[5] A. Bell, T. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, no. 6, pp. 1004-1034, 1995.

[6] A. Cichocki, S. Amari, K. Siwek et al., “ICALAB for Signal Processing - benchmarks”, *http://www.bsp.brain.riken.go.jp/ICALAB*.

[7] A. Feuerverger, “A consistent test for bivariate dependence,” *International Statistical Review*, vol. 61, no. 3, pp. 419-433, 1993.

[8] Timothy R. C. Read and Noel A. C. Cressie, *Goodness-of-fit statistics for discrete multivariate analysis*, Springer-Verlag, 1988