AN ADAPTIVE ALGORITHM FOR ICA

Joby Joseph and K. V. S. Hari

Indian Institute of Science, Bangalore, India, 560012

ABSTRACT

A new adaptive algorithm for Independent Component Analysis (ICA) has been developed. It directly applies stochastic gradient descent on the independence criterion. The cost surface is seen to be fairly smooth. The developed algorithm is always stable. It converges faster than earlier algorithm. The algorithm is developed for general $M \times M$ system case. Experiments are given to verify the above.

1. INTRODUCTION

Independent Component Analysis are methods by which statistically independent features are recovered from mixtures. In the current paper linear mixtures are considered. Given

$$\mathbf{x}(k) = \mathbf{H}\mathbf{s}(k) \tag{1}$$

where $\mathbf{s}(k) = [s_1(k)s_2(k)\ldots s_N(k)]$, $\mathbf{x}(k) = [x_1(k)x_2(k)\ldots x_N(k)]$. $\mathbf{H}$ is the mixing matrix. Problem is the estimation of the mixing system $\mathbf{H}$ and the original signal $\mathbf{s}(k)$. Principal Component Analysis (PCA) is the first step in the signal recovery [1]. It is assumed that PCA has been done for the original problem of signal separation. Thus $\mathbf{s}(k)$ have independent elements and a pre-whitening matrix $\mathbf{P}$ has been estimated such that

$$\mathbf{v}(k) = \mathbf{P}\mathbf{x}(k) \tag{2}$$

$$E\{\mathbf{v}(k)\mathbf{v}^T(k)\} = \mathbf{I} \tag{3}$$

Thus after the PCA stage

$$\mathbf{v}(k) = \mathbf{Q}\mathbf{s}(k) \tag{4}$$

$\mathbf{Q}$ is unitary and has to be estimated in order to recover $\mathbf{s}(k)$. Indirectly it can be stated as estimation of $\mathbf{W}$ unitary such that elements of $\mathbf{y}(k)$ are independent in

$$\mathbf{y}(k) = \mathbf{W}\mathbf{v}(k) \tag{5}$$

A good cost function to estimate $\mathbf{W}$ such that $\mathbf{y}$ has independent components, is the minimization of mutual information of elements of $\mathbf{y}$. $\mathbf{y}(k)$ is an estimate of $\mathbf{s}(k)$, but with ambiguity in the permutation of the elements. Under the whiteness constraint this is same as minimizing the entropies of the elements of $\mathbf{y}$.

$$\min_{\mathbf{W}} \sum_i \int f_{s_i}(y_i) \log(f_{s_i}(y_i)) dy_i \tag{6}$$

Here f_{s_i} is the density function of s_i. Here we deal with only s_i identically distributed. If the s_i are differently distributed then it is the signature used for identifying the recovered sources. This cost function can be used for estimating the sources within scaling and permutation if the sources are not Gaussian. This much is evident from the works [2],[3],[1]. There exists certain adaptive algorithms which estimate $\mathbf{W}$. Some which assume the source densities to be known, in which case the estimates converge faster. In some other attempts these densities are assumed to be of certain form, or the densities themselves are also estimated during the process. These are reviewed in [2]. Here we initially assume the source densities to be known and develop the adaptive algorithm.

2. ADAPTIVE ALGORITHM

We have to find $\mathbf{W}$ unitary such that (5) is satisfied. Thus $\mathbf{W}$ is a rotation matrix. In adaptive algorithm we start with $\mathbf{W}(0)$ the initial value and adapt it, using samples $\mathbf{y}(k)$, till it converges to the optimal $\mathbf{W}$. Thus this is a formulation of stochastic gradient descent algorithm. We need to find a sequence of $\mathbf{F}(k)$ such that $\mathbf{W}(k)$ converge to optimal $\mathbf{W}$.

$$\mathbf{W}(k+1) = \mathbf{F}(k+1)\mathbf{W}(k) \tag{7}$$

$$\mathbf{y}(k+1) = \mathbf{F}(k+1)\mathbf{W}(k)\mathbf{v}(k+1) \tag{8}$$

Pairwise rotation of the vectors forming $\mathbf{W}(k)$ is enough to achieve this [1]. So we develop an algorithm which will sequentially operate rotations on each pair so that the cost (6) is minimized. $\mathbf{W}$ and $\mathbf{F}$ are of size $N \times N$. So in an iteration where l^{th} and p^{th} elements of $\mathbf{y}$ forms the pair for processing, $\mathbf{F}(k)$ is of the form below where $\cos$ and $\sin$

terms are in the lth and p^{th}, columns and rows.

$$\mathbf{F}(k+1) = \begin{bmatrix} \mathbf{I} & & & \mathbf{0} \\ 0\ldots\ \cos(\theta(k)) & 0\ldots 1\, 0\ldots & \sin(\theta(k)) & 0\ldots \\ \mathbf{0} & \mathbf{I} & & \mathbf{0} \\ 0\ldots -\sin(\theta(k)) & 0\ldots 1\, 0\ldots & \cos(\theta(k)) & 0\ldots \\ \mathbf{0} & & & \mathbf{I} \end{bmatrix} \tag{9}$$

Thus the problem is reduced to finding a sequence $\theta_{lp}(1)$ $\theta_{lp}(2)\ \ldots \theta_{lp}(k)$ for each pair $\{\mathbf{y}_l(k), \mathbf{y}_p(k)\}$ such that, $\sum_0^k \theta_{lp}(k)$ tends to optimal ${}_{\mathbf{opt}}\theta_{lp}$. Now we make use of the ergodicity and stationarity of $\mathbf{s}(k)$ and do adaptation with instantaneous optimal ${}_{\mathbf{opt}}\theta_{lp}(k)$, where optimality is with respect to $\sum_i \log(f(y_i(k)))$ at instant k, so that the expected value in (6) is reached as $k \to \infty$. Instantaneous optimal rotation ${}_{\mathbf{opt}}\theta_{lp}(k)$ is given by solution to

$$\frac{d}{d\ {}_{\mathbf{opt}}\theta_{lp}(k)} \sum_i \log(f_{s_i}(y_i(k))) = 0 \tag{10}$$

$$\sum_{i\in\{l,p\}} \alpha_i(k) \frac{d\, y_i(k)}{d\ {}_{\mathbf{opt}}\theta_{lp}(k)} = 0 \tag{11}$$

where

$$\alpha_i(k) = \frac{f'(y_i(k))}{f(y_i(k))} \tag{12}$$

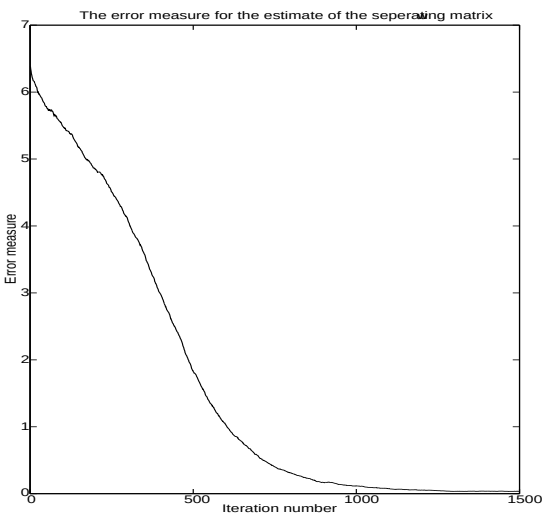

Fig. 1. The error with sample number is plotted in figure.

Notice that (10) arises from (11) because in (10) the terms other than $i \in \{l, p\}$ are zero when the pair $\{y_l, y_p\}$ is being considered. Substituting (9) and (8) in (11)

$$\alpha_l(k)(-y_l(k)\sin(\theta) + y_p(k)\cos(\theta)) + \alpha_p(k)(-y_l(k)\cos(\theta) \\ -y_p(k)\sin(\theta))|_{\theta={}_{\mathbf{opt}}\theta} = 0 \tag{13}$$

Solving for ${}_{\mathbf{opt}}\theta_{lp}(k)$,

$${}_{\mathbf{opt}}\theta_{lp}(k) = \tan^{-1}\left(\frac{\alpha_l(k)y_p(k) - \alpha_p(k)y_l(k)}{\alpha_p(k)y_p(k) + \alpha_l(k)y_l(k)}\right) \tag{14}$$

$$= \tan^{-1}(\beta(k)) \tag{15}$$

Approximating piecewise linearly and denoting sign of $\beta(k)$ with $sgn(\beta(k))$

$${}_{\mathbf{opt}}\theta_{lp}(k) \approx \begin{cases} \beta(k)\ if\ abs(\beta(k)) < \pi/2 \\ sgn(\beta(k))\pi/2\ if\ abs(\beta(k)) \geq \pi/2 \end{cases} \tag{16}$$

Hence the adaptation equation can be written as (8) and (9), with $\theta(k) = -\mu\ {}_{\mathbf{opt}}\theta(k)$, where $\mu < 1$ is the step size parameter.

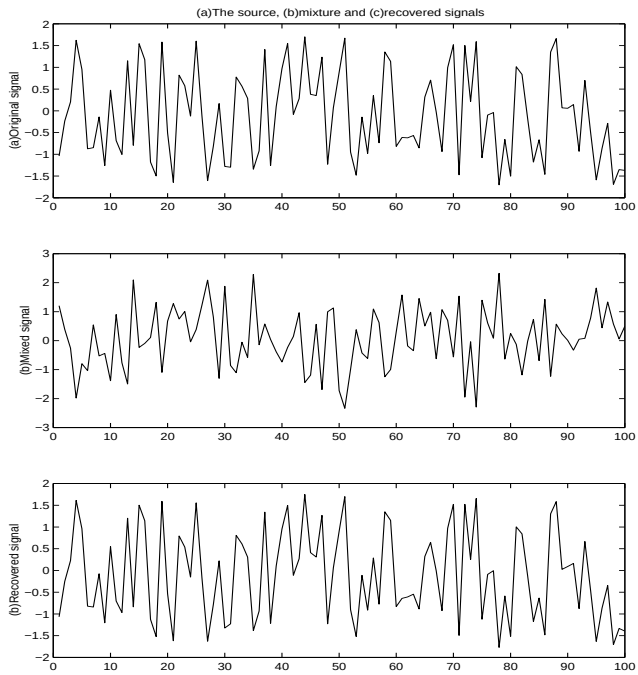

Fig. 2. A set of the source, the mixture and the recovered signals are plotted in this figure.

Algorithm

1. Initialize $\mathbf{W}(0)$.
2. Calculate $\mathbf{y}(k) = \mathbf{W}(k)\mathbf{v}(k)$.
3. Obtain ${}_{\mathbf{opt}}\theta(k)$ as in (16).
4. Form $\mathbf{F}(k)$ as in (9) with $\theta(k) = -\mu\ {}_{\mathbf{opt}}\theta(k)$
5. Update $\mathbf{W}(k)$ using (8)
6. Do steps 3 and 4 for each l, p pair.
7. Increment k and go to step (2)

3. STABILITY AND CONVERGENCE

Since $\mathbf{F}(k)$ is always unitary, the algorithm is always stable. As in general stochastic gradient descent algorithms, the solution will converge if a minima exist. This may be a local minima. To get an idea of the topology of the cost function, it can be plotted for some typical 2×2 cases. This can be done by estimating (6) from recovered sources, with $\mathbf{W}$ varying over different rotation angles. To do this choose $\mathbf{Q}$ of the form

$$\mathbf{Q} = \begin{bmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{bmatrix} \tag{17}$$

Then generate $\mathbf{s}(k)$ of required distribution and in turn obtain $\mathbf{v}(k)$, with (4). Generate $\mathbf{W}$ of the same form as $\mathbf{Q}$ in (17), but with angle ψ instead of ϕ. Vary ψ from $-\pi/2$ to $\pi/2$ and compute $\mathbf{y}(k)$ as in (5) for each ϕ . Estimate of (6) can be done from these samples of $\mathbf{y}(k)$ by finding its normalized histogram and using it as the estimate of probability moment function. This probability moment function being the sampled version of the density function $f_{y_i}()$.

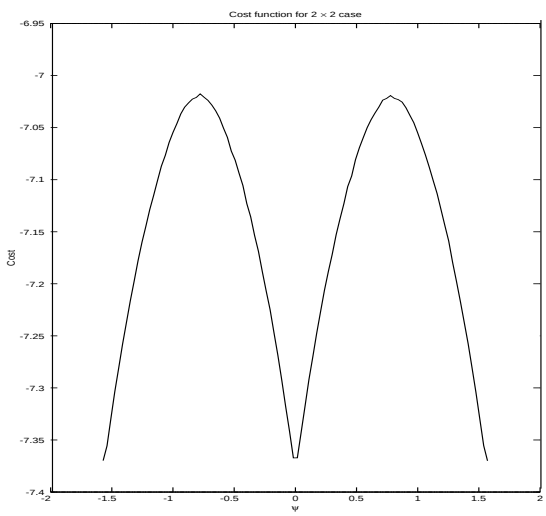

Fig. 3. The cost function for the 2×2 system $\mathbf{Q} = \mathbf{I}$

4. SIMULATIONS

Example 1. This example demonstrates the convergence of the algorithm for a 4×4 system. s_i are independent and uniformly distributed sources with unit variance. The 4×4 unitary mixing matrix $\mathbf{Q}$ was

$$\mathbf{Q} = \begin{bmatrix} -0.5793 & -0.1962 & -0.6218 & 0.4892 \\ 0.2594 & -0.5428 & -0.5362 & -0.5921 \\ 0.5601 & 0.5875 & -0.5477 & 0.2027 \\ -0.5324 & 0.5672 & -0.1609 & -0.6075 \end{bmatrix} \tag{18}$$

The error measure for the estimated $\mathbf{W}$ chosen is as described in ([1]). It checks how close $\mathbf{WQ}$ is to a permutation matrix. For the given case the error with sample number is plotted in figure(1). The curve is ensemble average of 100 runs. The step size parameter was 0.07. Also given in figure(2) is an example of recovered signal compared to the original and the mixed signals.

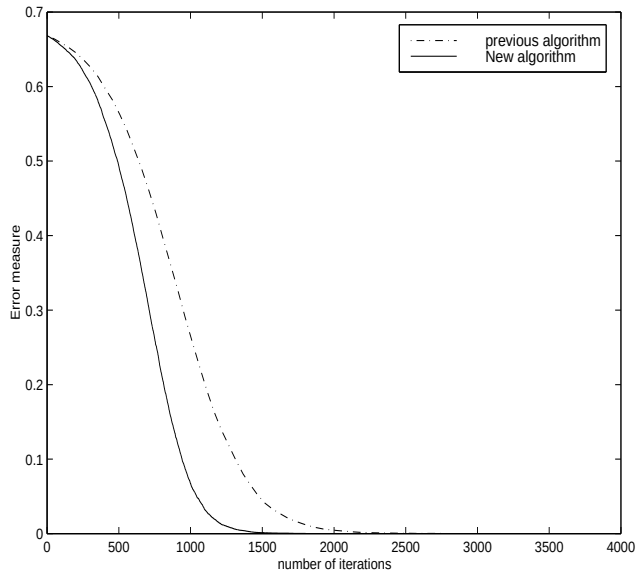

Fig. 4. The rate of convergence of the present algorithm is compared with that of earlier algorithm in this figure.

Example 2 In this example the cost surface for the stochastic gradient algorithm when the conditions are described as in (3). For the 2×2 system $\mathbf{Q} = \mathbf{I}$ the cost surface is plotted in figure (3), when the source density is uniform between $-\sqrt{3}$ and $\sqrt{3}$. If this density function is sampled to same number of points as the bins in the histogram used for computation from samples, and calculate entropy of this probability moment function, it is -7.38. Here the plotted cost is an average over 100 runs for each ψ. 100 equally spaced values of ψ between $-\pi/2$ and $\pi/2$ were chosen. Notice that there are two local minima with identical cost values. As seen from the figure (3) these minimum values correspond to the cost of the original unmixed signals, implying that both these minima to be equally valid solutions for the separation matrix $\mathbf{W}$.

Example 3 In this example the convergence of this new algorithm is compared with that in [3]. Identical 2×2 mix-

ing matrices were used and sources were uniform with unit variance. For new algorithm $\mu = 0.07$ and for the old algorithm $\beta = 0.002$ [3]. The convergence curves are given in figure (4). These plots are averaged over 100 runs of the algorithm. It is observed that the new algorithm converges faster to the same steady state error.

5. CONCLUSION

If the source density functions are known then this adaptive algorithm is good for recovery in the stage after PCA. The algorithm is stable irrespective of the parameters. It is optimizing an optimal cost function for ICA. The convergence is faster than the earlier algorithms. But computationally this is more expensive. As further work we are looking in to dynamic estimation of density function to avoid the requirement of apriori knowledge of the density function. Initial tests show that it is possible, but convergence is poor.

6. REFERENCES

[1] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[2] Jean-François Cardoso, "Blind signal separation: Statistical principles," *Proceedings of IEEE*, vol. 86, no. 10, pp. 2009–2025, October 1998.

[3] S. C. Douglas and S. Y. Kung, "Design of estimation/deflation approaches to independent component analysis," *Asilomar Conference*, pp. 707–711, 1998.