Blind Source Separation by Adaptive Estimation of Score Function Difference

Samareh Samadi[1], Massoud Babaie-Zadeh[1,2], Christian Jutten[3], and Kambiz Nayebi[1,⋆]

[1] Electrical Engineering Department
Sharif University of Technology, Tehran, Iran
{samarehsamadi,mbzadeh}@yahoo.com, knayebi@sina.sharif.edu
[2] Multimedia Lab. Iran Telecom Research Center(ITRC), Tehran, Iran
[3] Institut National Polytechnique de Grenoble (INPG)
Laboratoire des images et des signaux (LIS), Grenoble, France
Christian.Jutten@infg.fr

Abstract. In this paper, an adaptive algorithm for blind source separation in linear instantaneous mixtures is proposed, and it is shown to be the optimum version of the EASI algorithm. The algorithm is based on minimization of mutual information of outputs. This minimization is done using adaptive estimation of a recently proposed non-parametric "gradient" for mutual information.

1 Introduction

Blind Source Separation (BSS) is a relatively new subject in signal processing, which has been considered extensively since mid 80's [1]. It consists in retrieving unobserved independent mixed signals from mixtures of them, assuming there is information neither about the original sources, nor about the mixing system. The simplest BSS model is the linear instantaneous model. In this case, the mixture is supposed to be of the form $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{s}$ is the source vector, $\mathbf{x}$ is the observation vector, and $\mathbf{A}$ is the (constant) mixing matrix which is supposed to be an unknown matrix of full rank. The separating system, $\mathbf{B}$, tries to estimate the sources via $\mathbf{y} = \mathbf{B}\mathbf{x}$. For linear mixtures, it can be shown that the independence of the components of $\mathbf{y}$, is a necessary and sufficient condition for achieving the separation up to a scale and a permutation indeterminacy, provided that there is at most one Gaussian source [2].

The early works on BSS were concerned linear instantaneous mixture and by now a lot of algorithms are available for separating them (see [1, 3] for a review and extensive references. These methods can not be easily generalized to more complicated models. Source separation can be obtained by optimizing a "contrast function" *i.e.* a scalar measure of some "distributional property" of the outputs [4]. One of the most general contrast functions is mutual information,

⋆ This work has been partially funded by the European project Blind Source Separation and applications (BLISS, IST 1999-13077), Sharif University of Technology, and by Iran Telecom Research Center (ITRC).

C.G. Puntonet and A. Prieto (Eds.): ICA 2004, LNCS 3195, pp. 9–17, 2004.
© Springer-Verlag Berlin Heidelberg 2004

which has been shown [4] to be an asymptotically Maximum-Likelihood (ML) estimation of source signals. Recently a non-parametric "gradient" for mutual information has been proposed [5], which has been used successfully in separating different mixing models [6]. The proposed algorithms based on this gradient are all batch algorithms, which makes them unsuitable for being used in real-time applications.

In this paper, we propose an adaptive method for estimating this "gradient" of mutual information, and we use it to construct a new adaptive algorithm for separating linear instantaneous mixtures. This approach not only leads to good separation results, but also constructs a framework that can also be generalized to more complicated models. More interestingly, we will show that, for linear instantaneous mixtures, this new approach has a close relation to the famous EASI algorithm [7], and is, in fact, an optimal version of EASI. The paper is organized as follows. Section 2 reviews the essential materials to express the "gradient" of mutual information. The iterative equations of the algorithm are developed in Section 3, and its relation to EASI is considered in Section 4. The proposed algorithm is introduced as an optimum verion of EASI in section 5. This algorithm can be adaptively implemented using the adaptive estimation method of Section 6. In Section 7, the normalization method of the output energies is explained. Finally, Section 8 presents some experimental results.

2 Preliminary Issues

The objective of this section is to review mutual information definition, as the independence criterion, and its "gradient". Expressing this gradient, requires reviewing the definition of multivariate score functions of a random vector, which have been first introduced in [8].

2.1 Multivariate Score Functions

In statistics, the score function of a random variable y is defined as $-p'_y(y)/p_y(y)$, where $p_y(y)$ is the probability density function (PDF) of y. For an N-dimensional random vector $\mathbf{y} = (y_1, \ldots, y_N)^T$, two forms of score function are defined in [8]:

Definition 1 (MSF) *The marginal score function (MSF) of* $\mathbf{y}$*, is the vector of score functions of its components, i.e.:*

$$\boldsymbol{\psi}_{\mathbf{y}}(\mathbf{y}) = (\psi_1(y_1), \ldots, \psi_N(y_N))^T \tag{1}$$

where:

$$\psi_i(\mathbf{y}) = -\frac{d}{dy_i} \ln p_{y_i}(y_i) = -\frac{p'_{y_i}(y_i)}{p_{y_i}(y_i)} \tag{2}$$

and $p_{y_i}(y_i)$ *is the marginal PDF of* y_i*.*

Definition 2 (JSF) *The joint score function (JSF) of* $\mathbf{y}$*, is the vector function* $\boldsymbol{\varphi}_{\mathbf{y}}(\mathbf{y})$*, such that its* i*-th component is:*

$$\varphi_i(\mathbf{y}) = -\frac{\partial}{\partial y_i} \ln p_{\mathbf{y}}(\mathbf{y}) = -\frac{\frac{\partial}{\partial y_i} p_{\mathbf{y}}(\mathbf{y})}{p_{\mathbf{y}}(\mathbf{y})} \tag{3}$$

where $p_{\mathbf{y}}(\mathbf{y})$ *is the joint PDF of* $\mathbf{y}$*.*

Definition 3 (SFD) *The score function difference (SFD) of* $\mathbf{y}$*, is the difference between its JSF and MSF, i.e.:*

$$\boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y}) = \boldsymbol{\psi}_{\mathbf{y}}(\mathbf{y}) - \boldsymbol{\varphi}_{\mathbf{y}}(\mathbf{y}) \tag{4}$$

2.2 Mutual Information and Its Gradient

For measuring the statistical independence of random variables $y_1, \ldots, y_N$, one can use their mutual information, defined by:

$$\begin{aligned} I(\mathbf{y}) &= D\left(p_{\mathbf{y}}(\mathbf{y}) \parallel \prod_i p_{y_i}(y_i)\right) \\ &= \int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) \ln \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_i p_{y_i}(y_i)} d\mathbf{y} \\ &= E\left\{\ln \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_i p_{y_i}(y_i)}\right\} \end{aligned} \tag{5}$$

where $\mathbf{y} = (y_1, \ldots, y_N)^T$, and D denotes the Kullback-Leibler divergence. This function is always positive, and is zero if and only if the y_i's are independent.

For designing a source separation algorithm, one can use mutual information as a criterion for measuring output independence. In other words, the parameters of the separating system must be computed in such a way that the mutual information of the outputs be minimized. For doing this, the gradient based algorithms may be used. To calculate the gradient of the output mutual information with respect to the parameters of the separating system, the following theorem [5] will be quite helpful.

Theorem 1 *Let* $\boldsymbol{\Delta}$ *be a 'small' random vector, with the same dimension as* $\mathbf{x}$*. Then:*

$$I(\mathbf{x} + \boldsymbol{\Delta}) - I(\mathbf{x}) = E\left\{\boldsymbol{\Delta}^T \boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x})\right\} + o(\boldsymbol{\Delta}) \tag{6}$$

where $o(\boldsymbol{\Delta})$ *denotes higher order terms in* $\boldsymbol{\Delta}$*.*

This theorem points out that SFD can be called the "stochastic gradient" of mutual information.

Remark. Equation (6) may be stated in the following form (which is similar to what is done in [9]):

$$I(\mathbf{x} + \mathcal{E}\mathbf{y}) - I(\mathbf{x}) = E\left\{(\mathcal{E}\mathbf{y})^T \boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x})\right\} + o(\mathcal{E}) \tag{7}$$

where $\mathbf{x}$ and $\mathbf{y}$ are bounded random vectors, $\mathcal{E}$ is a matrix with small entries, and $o(\mathcal{E})$ stands for a term that converges to zero faster than $\|\mathcal{E}\|$. This equation is mathematically more sophisticated, because in (6) the term 'small random vector' is somewhat ad-hoc. Conversely, (6) is simpler, and easier to be used in developing gradient based algorithms for optimizing a mutual information.

3 Estimating Equations

In linear instantaneous mixture, the separating system is:

$$\mathbf{y} = \mathbf{B}\mathbf{x} \tag{8}$$

and $\mathbf{B}$ must be computed to minimize $I(\mathbf{y})$, where I stands for mutual information. For calculating $\mathbf{B}$, the steepest descent algorithm may be applied:

$$\mathbf{B}_{n+1} = \mathbf{B}_n - \mu \left.\frac{\partial I}{\partial \mathbf{B}}\right|_{\mathbf{B}=\mathbf{B}_n} \tag{9}$$

where μ is a small positive constant. However, to design an equivariant algorithm [7], that is, an algorithm whose separation performance does not depend on the conditioning of the mixing matrix, one must use the serial (multiplicative) updating rule:

$$\mathbf{B}_{n+1} = \left(\mathbf{I} - \mu \left[\nabla_{\mathbf{B}} I\right]_{\mathbf{B}=\mathbf{B}_n}\right) \mathbf{B}_n \tag{10}$$

where $\mathbf{I}$ denotes the identity matrix, and $\nabla_{\mathbf{B}} I \triangleq \frac{\partial I}{\partial \mathbf{B}} \mathbf{B}^T$ is the relative (or natural) gradient [7, 10] of $I(\mathbf{y})$ with respect to $\mathbf{B}$.

Using theorem 1, $\nabla_{\mathbf{B}} I$ can be easily obtained [5] (although, for this simple linear instantaneous case, this gradient may be directly calculated):

$$\nabla_{\mathbf{B}} I = E\left\{\boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\right\} \tag{11}$$

By dropping the expectation operation, the stochastic version of (10) is obtained:

$$\mathbf{B}_{n+1} = \left(\mathbf{I} - \mu \boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\right) \mathbf{B}_n \tag{12}$$

For developing the above algorithm in adaptive form, adaptive estimation of SFD is required, which will be discussed in Section 6.

4 Relation to EASI

The EASI algorithm has been proposed by Cardoso and Laheld [7]. In developing this algorithm, they showed that if the separation is achieved by minimizing a contrast function $\phi(\mathbf{B}) = E\{f(\mathbf{y})\}$ with respect to $\mathbf{B}$, the performance of the following serial updating algorithm, is independent of the mixing matrix:

$$\mathbf{B}_{n+1} = \left(\mathbf{I} - \mu \nabla \phi(\mathbf{B}_n)\right)\mathbf{B}_n \tag{13}$$

where the relative gradient $\nabla\phi(\mathbf{B})$ is:

$$\nabla\phi(\mathbf{B}) = \nabla E\left\{f(\mathbf{y})\right\} = E\left\{f'(\mathbf{y})\mathbf{y}^T\right\} \tag{14}$$

Consequently, the stochastic version of (13) becomes:

$$\mathbf{B}_{n+1} = \left(\mathbf{I} - \mu g(\mathbf{y})\mathbf{y}^T\right) \mathbf{B}_n \tag{15}$$

where $g \triangleq f'$. Developing EASI is then continued by choosing a "component-wise" g, and implementing a pre-whitening stage in the above algorithm, which is required by some contrast functions. This makes the final EASI equation more complicated than (15).

Now, let the contrast function be the mutual information:

$$\phi(\mathbf{B}) = I(\mathbf{y}) = E\left\{\ln \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_i p_{y_i}(y_i)}\right\} \tag{16}$$

Comparing with (14), we have:

$$f(\mathbf{y}) = \ln \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_i p_{y_i}(y_i)} \tag{17}$$

Then, the relative gradient (14) becomes (11), and the algorithm updating rule is (12). In fact, it is a special case of (15), where the contrast function is the mutual information of the outputs, and g is the SFD of $\mathbf{y}$. However, contrary to the "standard" EASI, where g is a "component-wise" and fixed function, here $g(\mathbf{y}) = \boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y})$ is a multi-variate function and depends on the distribution of $\mathbf{y}$.

5 Optimum EASI

As mentioned in sectin 4, $\phi(\mathbf{B})$ is a contrast function in EASI. Recall now that minimizing mutual information of outputs for source separation tends asymptotically towards a Maximum Likelihood (ML) estimation of sources [4]. Consequently, the optimal (in ML sense) contrast function in the EASI algorithm is mutual information of outputs and hence, the algorithm (12) can be considered as an *optimal* version of EASI (in ML sense). In other words, we have shown that the optimum choice of the non-linearity ($g(\mathbf{y})$) in the EASI algorithm is not a fixed and component-wise non-linearity, it is a multi-variate function which depends on the output statistics.

Moreover, in the "standard" EASI, one must take into account the necessity of existence of a pre-whitening stage, and implementing it in the algorithm. This makes the final equation of EASI [7] more complicated than (15). However, when using mutual information contrast, no pre-whitening is required.

Finally, besides its performance (see Section 8), one great advantage of this new algorithm is that it can be generalized to more complicated mixtures. In fact, it is based on SFD, which has been successfully used in separating other mixtures (especially, post-nonlinear and convolutive) in batch algorithms [6].

We recall that these advantages are obtained at the expense of higher computational load: a multi-variate nonlinear function (SFD) has to be estimated, based on the output statistics.

6 Adaptive SFD Estimation

For estimating the MSF, one must simply estimate the score functions of its components. It can be seen that for a function f with continuous first derivative and bounded sources we have [11]:

$$E\{f(x)\psi_x(x)\} = E\{f'(x)\} \tag{18}$$

where ψ_x is the MSF of the random variable x. Now, let the score function ψ_x be modeled as a linear combination of some basis functions $k_1(x)$, $k_2(x)$, ..., $k_L(x)$:

$$\hat{\psi}_x(x) = \sum_{i=1}^{L} w_i k_i(x) = \mathbf{k}(x)^T \mathbf{w} \tag{19}$$

where $\mathbf{k}(x) \triangleq (k_1(x), \ldots, k_L(x))^T$ and $\mathbf{w} \triangleq (w_1, \ldots, w_L)^T$. For calculating $\mathbf{w}$, we minimize the mean square error:

$$\mathcal{E} \triangleq E\left\{\left(\psi_x(x) - \hat{\psi}_x(x)\right)^2\right\} \tag{20}$$

Expanding the above expression and using (18), it is seen that the minimizer of $\mathcal{E}$ minimizes also:

$$\xi \triangleq \frac{1}{2}E\left\{\hat{\psi}_x(x)^2\right\} - E\left\{\frac{\partial}{\partial x}\hat{\psi}_x(x)\right\} \tag{21}$$

For minimizing ξ with respect to $\mathbf{w}$, the Newton method can be used:

$$\mathbf{w} \leftarrow \mathbf{w} - \mu E\left\{\left(\frac{\partial^2 \xi}{\partial \mathbf{w}^2}\right)\right\}^{-1} E\left\{\left(\frac{\partial \xi}{\partial \mathbf{w}}\right)\right\} \tag{22}$$

where:

$$\frac{\partial \xi}{\partial \mathbf{w}} = \mathbf{k}(x)\mathbf{k}(x)^T\mathbf{w} - \frac{\partial \mathbf{k}(x)}{\partial x} \tag{23}$$

and:

$$\frac{\partial^2 \xi}{\partial \mathbf{w}^2} = \mathbf{k}(x)\mathbf{k}(x)^T \tag{24}$$

This method can be easily generalized for estimating JSF. It has been shown [8] that for bounded sources and an arbitrary multivariate function $f(\mathbf{x})$ with continuous derivative with respect to x_i:

$$E\{f(\mathbf{x})\varphi_i(\mathbf{x})\} = E\left\{\frac{\partial}{\partial x_i} f(\mathbf{x})\right\} \tag{25}$$

Let now $\varphi_i(\mathbf{x})$, the i-th component of JSF, be estimated as the linear combination of the (multivariate) basis functions $k_1(\mathbf{x}), ..., k_L(\mathbf{x})$, that is:

$$\hat{\varphi}_i(\mathbf{x}) = \sum_{i=1}^{L} w_i k_i(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T \mathbf{w} \tag{26}$$

and $\mathbf{w}$ is the minimizer of $E\left\{(\varphi_i(\mathbf{x}) - \hat{\varphi}_i(\mathbf{x}))^2\right\}$.

Following similar calculation as above, we obtain the same algorithm given by equations (22), (23) and (24), where in this case:

$$\xi = \frac{1}{2}E\left\{\hat{\varphi}_i(\mathbf{x})^2\right\} - E\left\{\frac{\partial}{\partial x_i}\hat{\varphi}_i(\mathbf{x})\right\} \tag{27}$$

Finally, SFD is estimated by calculating the difference of the estimated MSF and JSF.

7 Normalization of Output Energies

From the scale indeterminacy it is deducted that the algorithm (12) has no restriction on the energy of outputs. Consequently, this algorithm does not converge to a unique solution. To overcome this indeterminacy, and making the algorithm to converge to unit energy outputs, we replace the i-th diagonal element of $\boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T$ by $1-y_i^2$ to force the separating system to create unit variance outputs. This is similar to what is done in [11].

8 Experimental Result

As an experiment, two independent sources with normal and uniform distributions and with zero means and unit variances are mixed by:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.7 \\ 0.5 & 1 \end{bmatrix} \tag{28}$$

Basis functions for estimating $\psi_i(y_i)$ are:

$$k_1(y) = 1,\, k_2(y) = y,\, k_3(y) = y^2,\, k_4(y) = y^3 \tag{29}$$

and basis functions for estimating $\varphi_i(\mathbf{y})$ are:

$$\begin{aligned} &k_1(y_1, y_2) = 1, \\ &k_2(y_1, y_2) = y_1,\, k_3(y_1, y_2) = y_1^2,\, k_4(y_1, y_2) = y_1^3 \\ &k_5(y_1, y_2) = y_2,\, k_6(y_1, y_2) = y_2^2,\, k_7(y_1, y_2) = y_2^3 \end{aligned}$$

To compare the separation result of the proposed algorithm with EASI, we have separated this mixture, using both algorithms. In our method, the adaptation rate of the Newton algorithm is 0.1 and the adaptation rate of the separation algorithm is 0.001. In EASI, the component-wise nonlinear function $g(y_i) = y_i|y_i|^2$ has been used, with the same adaptation rate (0.001). Figure 1 shows the averaged output signal to noise ratios (SNR) taken over 50 runs of the algorithms. SNR is defined as:

$$\mathrm{SNR} = 10 \log_{10} \frac{E\left\{s^2\right\}}{E\left\{(y-s)^2\right\}} \tag{30}$$

where y is the output corresponding to the source s. The figure shows that the proposed algorithm has better separation performance than EASI, as was expected because this algorithm is an optimal version of EASI (see Section 4). However the cost of this better performance is a higher complexity (which increases with the source number), since a multivariate non-linear function must be estimated at each iteration.

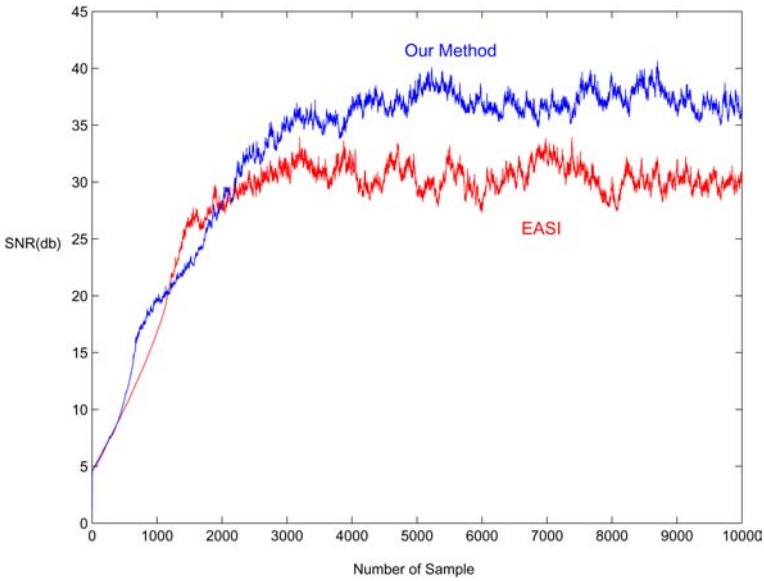

Fig. 1. Output SNRs versus iteration for EASI and our method.

9 Conclusion

In this paper an adaptive algorithm for blind separating linear instantaneous mixtures has been proposed, which is based on adaptive estimation of SFD. It has been shown that this algorithm can be seen as an optimum version of the EASI algorithm. Moreover, it is conjectured that this method can be generalized to separating more complicated (than linear instantaneous) mixing models, such as convolutive and non-linear mixtures. This is because SFD has been successfully used in separating these models [6]. Such a generalization is currently under study. The drawback of this method is that, despite of EASI, this algorithm requires the estimation of multivariate score functions (which are related to joint PDFs). This estimation becomes too difficult, and requires a lot of data, when the dimension (*i.e.* number of sources) grows. Practically, this method is suitable only up to 3 or 4 sources.

References

1. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiely & Sons, 2001.
2. P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
3. A. Cichocki and S.I. Amari, *Adaptive Blind Signal and Image Processing*, Wiley, 2002.
4. J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of IEEE*, vol. 9, pp. 2009–2025, 1998.

5. M. Babaie-Zadeh, C. Jutten, and K. Nayebi, "Differential of mutual information function," *IEEE Signal Processing Letters*, vol. 11, no. 1, pp. 48–51, January 2004.
6. M. Babaie-Zadeh, *On blind source separation in convolutive and nonlinear mixtures*, Ph.D. thesis, INP Grenoble, 2002.
7. J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on SP*, vol. 44, no. 12, pp. 3017–3030, December 1996.
8. M. Babaie-Zadeh, C. Jutten, and K. Nayebi, "Separating convolutive mixtures by mutual information minimization," in *Proceedings of IWANN'2001*, Granada, Spain, Juin 2001, pp. 834–842.
9. D. T. Pham, "Mutual information approach to blind separation of stationary sources," *IEEE Transactions on Information Theory*, vol. 48, no. 7, pp. 1–12, July 2002.
10. S. I. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, 1998.
11. A. Taleb and C. Jutten, "Entropy optimization, application to blind source separation," in *ICANN*, Lausanne, Switzeland, October 1997, pp. 529–534.