Separating convolutive mixtures by mutual information minimization

Massoud Babaie-Zadeh[1,2], Christian Jutten[1], and Kambiz Nayebi[2]

[1] Institut National Polytéchnique de Grenoble (INPG), Laboratoire des Images et des Signaux (LIS), 46 Avenue Félix Viallet, Grenoble, France
[2] Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran

Abstract. Blind Source Separation (BSS) is a basic problem in signal processing. In this paper, we present a new method for separating convolutive mixtures based on the minimization of the output mutual information. We also introduce the concept of joint score function, and derive its relationship with marginal score function and independence. The new approach for minimizing the mutual information is very efficient, although limited by multivariate distribution estimations.

1 Introduction

Blind Source Separation (BSS) is a basic problem in signal processing, which has been considered intensively in the last decade. In the linear instantaneous 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. The separator system, $\mathbf{B}$, tries to estimate the sources via $\mathbf{y} = \mathbf{B}\mathbf{x}$. For linear mixtures, it can be shown that 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) [3].

The convolutive case, too, has been addressed by a few authors [9, 8, 4, 5, 2, 6]. In that case, the mixing and separating matrices are linear time invariant (LTI) filters, *i.e.* the mixing system is:

$$\mathbf{x}(n) = [\mathbf{A}(z)]\,\mathbf{s}(n) \tag{1}$$

and the separating system is:

$$\mathbf{y}(n) = [\mathbf{B}(z)]\,\mathbf{x}(n) \tag{2}$$

For these mixtures too, it has been shown that the independence of the outputs is necessary and sufficient for signal separation (up to a filtering and a permutation indeterminacy) [9]. However, in convolutive mixtures, the independence of two random processes y_1 and y_2, cannot be deduced from the sole independence of $y_1(n)$ and $y_2(n)$, but requires the independence of $y_1(n)$ and $y_2(n-m)$, for all n and all m.

Several methods have been proposed for satisfying the above condition. Most of them are based on higher (than 2) order statistics: cancellation of cross-spectra

[9], of higher order cross-moments [8], of higher order cross-cumulants [8, 2], or more generally on a contrast function [6].

In this paper, we introduce a method based on minimizing the output mutual information. This method is inspired by the method proposed by Taleb and Jutten [7] for the instantaneous mixtures, but contains a few new points. Section 2 contains preliminary results on stochastic process independence and the definition and a few properties of joint score function. Estimating equations are derived in Section 3. The algorithm and experiments are presented in Section 4 and 5, respectively.

2 Preliminary Issues

2.1 Independence in the Convolutive Context

In convolutive mixtures, $y_1(n)$ and $y_2(n)$ can be independent, while y_1 and y_2 are not[1]. For example, if the sources s_i are iid, and:

$$\mathbf{B}(z)\mathbf{A}(z) = \begin{bmatrix} 1 & z^{-1} \\ 0 & 1 \end{bmatrix} \tag{3}$$

then the outputs are:

$$\begin{cases} y_1(n) = s_1(n) + s_2(n-1) \\ y_2(n) = s_2(n) \end{cases} \tag{4}$$

It is obvious that in this case, $y_1(n)$ and $y_2(n)$ are independent for all n, but y_1 and y_2 are not, and thus the source separation is not achieved.

To check the independence of two random variables x and y, one can use the mutual information:

$$I(x, y) = \int_{x,y} p_{xy}(x, y) \ln \frac{p_{xy}(x, y)}{p_x(x) p_y(y)} dx dy = \textstyle\sum_i H(y_i) - H(\mathbf{y}) \tag{5}$$

where H denotes the Shannon entropy. $I(x, y)$ is always non-negative, and vanishes if and only if the random variables x and y are independent.

However, $I(y_1(n), y_2(n)) = 0$ cannot be a separation criterion. Conversely, one can use $I(y_1(n), y_2(n-m)) = 0$ for all m. But, this criterion, for all m is practically untractable. Thus, we restrict ourselves to a finite set, say $m \in \{-M, \ldots, +M\}$. For example, Charkani [2] and Nguyen and Jutten [8] considered the independence of $y_1(n)$ and $y_2(n-m)$ for $m \in \{0, \ldots, +M\}$, where M is the maximum degree of the FIR filters of their separating structure.

[1] Recall that, by definition, two stochastic processes X_1 and X_2 are independent if and only if the random variables $X_1(n)$ and $X_2(n-m)$ are independent for all n and all m.

2.2 JSFs versus MSFs

In this subsection, we introduce the concepts of Joint Score Function (JSF) and Marginal Score Function (MSF).

Definition 1 (Score Function) *Score function of a scalar random variable x, is the log derivative of its density, i.e.:*

$$\psi(x) = \frac{d}{dx} \ln p_x(x) = \frac{p'_x(x)}{p_x(x)} \tag{6}$$

where p_x denotes the Probability Density Function (PDF) of x.

For the N dimensional random vector $\mathbf{x} = (x_1, \ldots, x_N)^T$, we define two score functions:

Definition 2 (MSF) *Marginal Score Function (MSF) of $\mathbf{x}$, is the vector of score functions of its components, i.e.:*

$$\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x}) = (\psi_{x_1}(x_1), \ldots, \psi_{x_N}(x_N))^T \tag{7}$$

Note that the ith component of $\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x})$ is:

$$\psi_i(\mathbf{x}) = \frac{d}{dx_i} \ln p_{x_i}(x_i) \tag{8}$$

where $p_{x_i}(x_i)$ is the PDF of x_i.

Definition 3 (JSF) *Joint Score Function (JSF) of $\mathbf{x}$, is the vector function $\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x})$, such that its ith component is:*

$$\varphi_i(\mathbf{x}) = \frac{\partial}{\partial x_i} \ln p_{\mathbf{x}}(\mathbf{x}) = \frac{\frac{\partial}{\partial x_i} p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{x}}(\mathbf{x})} \tag{9}$$

where $p_{\mathbf{x}}(\mathbf{x})$ is the mutual PDF of $\mathbf{x}$.

Generally, MSF and JSF are not equal, but we have the following theorem:

Theorem 1 *The components of the random vector $\mathbf{x}$ are independent if and only if its JSF and MSF are equal, i.e.:*

$$\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x}) = \boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x}) \tag{10}$$

For a proof, refer to appendix.

Definition 4 (SFD) *Score Function Difference (SFD) of $\mathbf{x}$ is the difference between its JSF and MSF, i.e.:*

$$\boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x}) = \boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x}) - \boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x}) \tag{11}$$

As a consequence of Theorem 1, SFD is an independence criterion.

3 Estimating Equations

3.1 Estimating MSF and JSF

For estimating the MSF, one must simply estimate the score functions of its components. In [7], the following theorem is used for estimating the score function of a scalar random variable:

Theorem 2 *Consider a scalar random variable x, and a function f with a continuous first derivative, satisfying:*

$$\lim_{x \to \pm\infty} f(x) p_x(x) = 0 \tag{12}$$

then :

$$E\{f(x)\psi(x)\} = -E\{f'(x)\} \tag{13}$$

Note that the condition (12) is not very restrictive, since most of densities $p_x(x)$ vanishes as x tends towards infinity.

Now, consider the score function estimate equal to a linear combination of some kernel functions $k_i(x)$, *i.e.*:

$$\hat{\psi}(x) = \sum_{i=1}^{L} w_i k_i(x) = \mathbf{k}^T(x)\mathbf{w} \tag{14}$$

where $\mathbf{k}(x) = (k_1(x), \ldots, k_L(x))^T$ and $\mathbf{w} = (w_1, \ldots, w_L)^T$. We estimate $\mathbf{w}$ for minimizing the mean square error $E\left\{[\psi(x) - \hat{\psi}(x)]^2\right\}$. Applying the orthogonality principle, and using Theorem 2, $\mathbf{w}$ is obtained by:

$$E\left\{\mathbf{k}(x)\mathbf{k}^T(x)\right\}\mathbf{w} = E\{\mathbf{k}(x)\psi(x)\} = -E\{\mathbf{k}'(x)\} \tag{15}$$

This method can be easily generalized to multivariate PDF. First, we prove the generalization of Theorem 2:

Theorem 3 *Consider a random vector $\mathbf{x} = (x_1, \ldots, x_N)^T$, and a multivariate scalar function f with continuous derivatives with respect to x_i, satisfying:*

$$\lim_{x_i \to \pm\infty} \int_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_N} f(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) dx_1 \cdots dx_{i-1} dx_{i+1} dx_N = 0 \tag{16}$$

then:

$$E\{f(\mathbf{x})\varphi_i(\mathbf{x})\} = -E\left\{\frac{\partial}{\partial x_i} f(\mathbf{x})\right\} \tag{17}$$

For a proof, refer to appendix.

Now, suppose we model $\varphi_i(\mathbf{x})$, the ith component of JSF, as a linear combination of the kernel functions $k_1(\mathbf{x}), \ldots, k_L(\mathbf{x})$, *i.e.* $\hat{\varphi}_i(\mathbf{x}) = \mathbf{k}^T(\mathbf{x})\mathbf{w}$, where $\mathbf{k}(\mathbf{x}) = (k_1(\mathbf{x}), \ldots, k_L(\mathbf{x}))^T$. Following similar computations than those used for developing (15), it can be shown:

$$E\left\{\mathbf{k}(\mathbf{x})\mathbf{k}^T(\mathbf{x})\right\}\mathbf{w} = -E\left\{\frac{\partial}{\partial x_i}\mathbf{k}(\mathbf{x})\right\} \tag{18}$$

3.2 Gradient of the Mutual Information

Suppose the separating system consists of FIR filters with a maximum degree of p. Hence, the separating system writes as:

$$\mathbf{B}(z) = \mathbf{B}_0 + \mathbf{B}_1 z^{-1} + \cdots + \mathbf{B}_p z^{-p} \tag{19}$$

For developing a gradient-based algorithm, we must estimate the derivative of the mutual information with respect to each matrix $\mathbf{B}_k$.

Theorem 4 *If the separating system $\mathbf{B}(z)$ satisfies (19), then:*

$$\frac{\partial}{\partial \mathbf{B}_k} I\left(y_1(n), y_2(n-m)\right) = E\left\{\boldsymbol{\beta}^{(m)}(n)\mathbf{x}^T(n-k)\right\} \tag{20}$$

where I denotes the mutual information, and $\boldsymbol{\beta}^{(m)}(n)$ is defined by:

$$\boldsymbol{\beta}(n) = \boldsymbol{\beta}_{y_1(n),y_2(n-m)}(y_1(n), y_2(n-m)) \tag{21}$$

$$\boldsymbol{\beta}^{(m)}(n) = \begin{bmatrix} \beta_1(n) \\ \beta_2(n+m) \end{bmatrix} \tag{22}$$

where $\beta_{\mathbf{x}}$ denotes the SFD of the random vector $\mathbf{x}$.

For a proof refer to appendix.

In other words, for computing $\boldsymbol{\beta}^{(m)}(n)$, one must first shift forward the second component of $\mathbf{y}$, then compute its SFD to obtain $\boldsymbol{\beta}(n)$, and then shift back its second component to obtain $\boldsymbol{\beta}^{(m)}(n)$.

Note that the algorithm stops when $\boldsymbol{\beta}(n) = 0$, which is equivalent to the independence of the outputs.

4 The Algorithm

The steepest descent algorithm has been used for achieving the output independence, *i.e.* in each iteration, all of the $\mathbf{B}_k$s are updated according to:

$$\mathbf{B}_k = \mathbf{B}_k - \mu \frac{\partial}{\partial \mathbf{B}_k} I(y_1(n), y_2(n-m)) \tag{23}$$

where μ is a small positive constant, the derivative is computed following (20), and SFD is computed using (14), (15) and (18). For estimating the MSFs, we have chosen the 4 kernels ($L = 4$):

$$k_1(x) = 1 \ , \ k_2(x) = x \ , \ k_3(x) = x^2 \ , \ k_4(x) = x^3 \tag{24}$$

For estimating the JSFs, we used the 7 kernels ($L = 7$):

$$\begin{aligned} k_1(x_1,x_2) = 1 \ , \ k_2(x_1,x_2) = x_1 \ , \ k_3(x_1,x_2) = x_1^2 \ , \ k_4(x_1,x_2) = x_1^3 \\ , \ k_5(x_1,x_2) = x_2 \ , \ k_6(x_1,x_2) = x_2^2 \ , \ k_7(x_1,x_2) = x_2^3 \end{aligned} \tag{25}$$

For tractability of the algorithm, the value of m is randomly chosen from the set $\{-M, \ldots, M\}$ at each iteration, implying a stochastic implementation of the independence criterion. Note that this algorithm is not equivariant [1], and consequently its performance is not independent of the mixing filter.

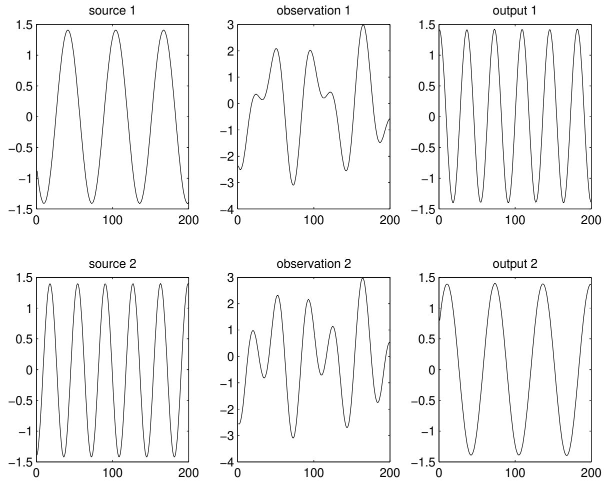

Fig. 1. Separating two sinusoids.

5 Experimental Results

For measuring the separation performance, we define the output SNR. Assuming no permutation, and denoting $\mathbf{C}(z) = \mathbf{B}(z)\mathbf{A}(z)$, the output SNR on the first channel is:

$$\mathrm{SNR}_1 = \frac{E\left\{[y_1(n)]^2\right\}}{E\left\{\{[C_{12}(z)]\, s_2(n)\}^2\right\}} \tag{26}$$

Hence a high SNR_1 means that the effect of the second source in the first output is negligible. However, the first output is not necessarily equal to the first source, it can be a filtered version of that (see Sect. 2.1).

In the first experiment, we have chosen two sinusoids (500 samples), with different frequencies, and mixed them with the $\mathbf{A}(z)$:

$$\begin{bmatrix} 1 + 0.2z^{-1} + 0.1z^{-2} & 0.5 + 0.3z^{-1} + 0.1z^{-2} \\ 0.5 + 0.3z^{-1} + 0.1z^{-2} & 1 + 0.2z^{-1} + 0.1z^{-2} \end{bmatrix} \tag{27}$$

We thus used a 2-degree FIR separating filter, $\mu = 0.3$, and $M = 4$.

Figure 1 shows the separation results, after 13000 iterations. The output SNRs are 48.64dB and 48.61dB. Figure 5, shows time variation of output SNRs.

In the second experiment, we mixed two uniform white noises, with zero means and unit variances. We choose the same mixing system and parameters as in the first experiment. Figure 5 shows the output SNRs in dB.

6 Conclusion

In this paper, we proposed a new method for separating the convolutive mixtures, based on a stochastic implementation of the minimization of delayed output mu-

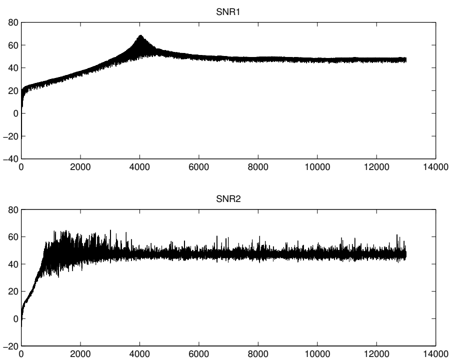

Fig. 2. Output SNRs, in separating two sinusoids.

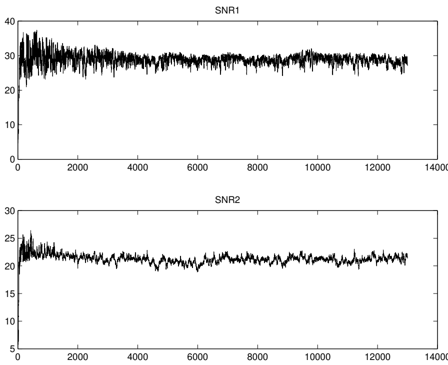

Fig. 3. Output SNRs, in separating two uniform white noises.

tual informations. Moreover, each mutual information term is minimized using Marginal and Joint score functions. The experiments show its efficiency. The main restriction of this new method is related to JSF estimation which requires large samples, and whose implementation will be difficult for more than 3 or 4 sources.

A Appendix

Proof of Theorem 1: The proof is given in the two dimensional case. Its generalization to higher dimensions is obvious.

If the components of $\mathbf{y}$ are independent, then (10) can be easily obtained. Conversely, suppose that (10) holds, then we prove that the components of $\mathbf{y}$ are independent. Following (10), we have $\frac{\partial}{\partial y_1} \ln p_{\mathbf{y}}(y_1, y_2) = \frac{\partial}{\partial y_1} \ln p_{y_1}(y_1)$. Integrating both sides of this equation with respect to y_1, leads to:

$$\ln p_{\mathbf{y}}(y_1, y_2) = \ln p_{y_1}(y_1) + g(y_2) \Rightarrow p_{\mathbf{y}}(y_1, y_2) = p_{y_1}(y_1) h(y_2) \tag{28}$$

By integrating both sides of this equation with respect to y_1 from $-\infty$ to $+\infty$, we have $h(y_2) = p_{y_2}(y_2)$ thus the result holds. $\square$

Proof of Theorem 3: Without loss of generality, let $i = 1$. We have:

$$\begin{aligned} E\left\{f(\mathbf{x})\varphi_1(\mathbf{x})\right\} &= \int f(\mathbf{x})\varphi_1(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})d\mathbf{x} \\ &= \int_{x_2,\ldots,x_N}\int_{x_1} f(\mathbf{x})\frac{\partial}{\partial x_1}p_{\mathbf{x}}(\mathbf{x})dx_1 dx_2 \cdots dx_N \end{aligned} \tag{29}$$

Applying integration by parts for the inner integral and using (17) leads to the desired relation. $\square$

Proof of Theorem 4: Here, because of the limited space, we only prove the theorem for $m = 0$. The generalization to the case $m \neq 0$ is straightforward, but contains somewhat complicated calculations.

Let $\mathbf{B}(z)$ satisfy (19), and $b_{ij}^{(k)}$ denote the ijth element of $\mathbf{B}_k$, then:

$$\frac{\partial H(\mathbf{y}(n))}{\partial b_{ij}^{(k)}} = -E\left\{\frac{\partial}{\partial b_{ij}^{(k)}} \ln p_{\mathbf{y}}(\mathbf{y})\right\} \tag{30}$$

Among the y_1 through y_N, only y_i depends on $b_{ij}^{(k)}$, hence:

$$\frac{\partial H(\mathbf{y}(n))}{\partial b_{ij}^{(k)}} = -E\left\{\frac{\partial \ln p_{\mathbf{y}(n)}(\mathbf{y}(n))}{\partial y_i(n)} \cdot \frac{\partial y_i(n)}{\partial b_{ij}^{(k)}}\right\} = -E\left\{\varphi_{\mathbf{y},i}(n) x_j(n-k)\right\} \tag{31}$$

Consequently:

$$\frac{\partial H(\mathbf{y}(n))}{\partial \mathbf{B}_k} = -E\left\{\boldsymbol{\varphi}_{\mathbf{y}}(n)\mathbf{x}^T(n-k)\right\} \tag{32}$$

We now compute the marginal entropy derivatives:

$$\begin{aligned}\frac{\partial}{\partial b_{ij}^{(k)}} \textstyle\sum_i H(y_i(n)) &= \frac{\partial}{\partial b_{ij}^{(k)}} H(y_i(n)) = -E\left\{\frac{\partial}{\partial b_{ij}^{(k)}} \ln p_{y_i(n)}(y_i(n))\right\} \\ &= -E\left\{\frac{\partial \ln p_{y_i(n)}(y_i(n))}{\partial y_i(n)} \cdot \frac{\partial y_i(n)}{\partial b_{ij}^{(k)}}\right\} = -E\left\{\psi_{y_i(n)}(n) x_j(n-k)\right\}\end{aligned} \tag{33}$$

Hence:

$$\frac{\partial}{\partial \mathbf{B}_k} \sum_i H(y_i) = -E\left\{\boldsymbol{\psi}_{\mathbf{y}}(n)\mathbf{x}^T(n-k)\right\} \tag{34}$$

Combining (5), (32) and (34) proves the theorem. □

References

[1] J.-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. on SP*, 44(12):3017–3030, December 1996.

[2] N. Charkani. *Séparation auto-adaptative de sources pour des mélanges convolutifs. Application à la téléphonie mains-libres dans les voitures.* Thèse de l'INP Grenoble, 1996.

[3] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[4] S.V. Gerven and D.V. Compernolle. Signal separation by symmetric adaptive decorrelation: stability, convergence and uniqueness. *IEEE trans. on SP*, 43(7):1602–1612, 1995.

[5] U. Lindgren and H. Broman. Source separation using a criterion based on second order statistics-an algebric approach. *IEEE trans. on SP*, 46(7):1837–1850, 1998.

[6] C. Simon. *Séparation aveugle des sources en mélange convolutif.* PhD thesis, l'université de Marne la Vallée, Novembre 1999. (In French).

[7] A. Taleb and C. Jutten. Entropy optimization, application to blind source separation. In *ICANN*, pages 529–534, Lausanne, Switzeland, October 1997.

[8] H.L. Nguyen Thi and C. Jutten. Blind sources separation for convolutive mixtures. *Signal Processing*, 45:209–229, 1995.

[9] D. Yellin and E. Weinstein. Criteria for multichannel signal separation. *IEEE Trans. Signal Processing*, pages 2158–2168, August 1994.