

Differential of the Mutual Information

Massoud Babaie-Zadeh, Christian Jutten, *Member, IEEE*, and Kambiz Nayebi

Abstract—In this letter, we compute the variation of the mutual information, resulting from a small variation in its argument. Although the result can be applied in many problems, we consider only one example: the result is used for deriving a new method for blind source separation in linear mixtures. The experimental results emphasize the performance of the resulting algorithm.

Index Terms—Blind source separation (BSS), independent component analysis (ICA), mutual information.

I. INTRODUCTION

BLIND SOURCE SEPARATION (BSS) and independent component analysis (ICA) are basic problems in signal processing that have been studied intensively in the last 15 years. For linear instantaneous mixtures, the problem is stated as follows. Let $\mathbf{s}(n) = (s_1(n), \dots, s_N(n))^T$ be the vector of some statistically independent source signals that are mixed by a regular mixing matrix \mathbf{A} and generate the observed signals $\mathbf{x}(n) = (x_1(n), \dots, x_N(n))^T$, i.e., $\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n)$ (in this letter, the number of sources and the number of observations are assumed to be equal). The goal of BSS is to retrieve the source signals s_i only by observing x_i 's: there is neither information about the source signals (but their statistical independence) nor about the mixing matrix \mathbf{A} . For separating the mixture, we estimate the separating matrix \mathbf{B} such that the components of the output vector $\mathbf{y} = \mathbf{B}\mathbf{x}$ become statistically independent. It has been proven [1] that if there is no more than one Gaussian source, and if the components of \mathbf{y} are independent, they will be a copy of the source signals up to a scaling and a permutation indeterminacy.

The degree of independence between random variables y_1, y_2, \dots, y_N can be measured by their mutual information

$$I(\mathbf{y}) = \int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) \ln \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_i p_{y_i}(y_i)} d\mathbf{y} = \sum_i H(y_i) - H(\mathbf{y}) \quad (1)$$

where $\mathbf{y} = (y_1, \dots, y_N)^T$, $p_{\mathbf{y}}$ and p_{y_i} are the probability density functions (pdfs) of \mathbf{y} and y_i , respectively, and H denotes Shannon's entropy. The mutual information is always nonnega-

tive and vanishes if and only if the random variables y_1, \dots, y_N are independent. Therefore, the estimation algorithm of the separating matrix \mathbf{B} can be designed based on minimizing the mutual information of the outputs $I(\mathbf{y})$, and for this minimization, the steepest descent algorithm can be used. This technique has already been applied successfully for separating linear instantaneous mixtures [2], [3] as well as nonlinear mixtures [4].

Usually, $I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{y})$ is modified by using the multiplicative relation

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{p_{\mathbf{x}}(\mathbf{x})}{|\det \mathbf{B}|} \quad (2)$$

which leads to $H(\mathbf{y}) = H(\mathbf{x}) + \ln |\det \mathbf{B}|$ and consequently

$$I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{x}) - \ln |\det \mathbf{B}|. \quad (3)$$

The gradient of $I(\mathbf{y})$ with respect to \mathbf{B} is then easily obtained and only requires estimation of marginal pdfs or more exactly of their log-derivatives. Similar relations, and the gradient of $I(\mathbf{y})$, can be derived for particular nonlinear mixtures [4].

However, for more complicated mixtures, such as convolutive mixtures (where the mixing matrix is composed of filters, instead of simple scalars), or convolutive-nonlinear mixtures, a simple multiplicative relation like (2) does not exist. In these cases, if we know the variation of the mutual information resulting from a small variation of its argument (the *differential* of mutual information), then we can easily design gradient-based algorithms. The main purpose of this letter is to calculate this *differential*. In this letter, we only apply it to source separation, but the result is very general and could be used in many domains where the mutual information gradient is required. The letter is organized as follows. In Section II, we introduce a few definitions. Section III is devoted to the computation of the differential of the mutual information. In Section IV, this result is applied for deriving estimation equations for linear instantaneous and convolutive mixtures. In Section V, we illustrate the algorithm efficacy by a simple experiment.

II. JSFS, MSFS, AND SDFS

In this section, we introduce the definition of the joint score function (JSF), the marginal score function (MSF), and the score function difference (SFD). First, recall the definition of the score function of a scalar random variable from statistics literature.

Definition 1: The score function of the scalar random variable x is the opposite of the log-derivative of its density, i.e.,

$$\psi(x) = -\frac{d}{dx} \ln p_x(x) = -\frac{p'_x(x)}{p_x(x)}. \quad (4)$$

Now, let $\mathbf{x} = (x_1, \dots, x_N)^T$ be an N -dimensional random vector. We then define two different forms of score functions.

Manuscript received May 16, 2002; revised December 9, 2002. This work was supported in part by the European project BLISS (IST-1999-14190). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. P. C. Ching.

M. Babaie-Zadeh is with the Laboratoire des Images et des Signaux (LIS), Institut National Polytechnique de Grenoble (INPG), 38031 Grenoble Cedex, France and also with the Electrical engineering Department, Sharif University of Technology, Tehran, Iran.

C. Jutten is with the Laboratoire des Images et des Signaux (LIS), Institut National Polytechnique de Grenoble (INPG), 38031 Grenoble Cedex, France and also with the Institut des Sciences et Techniques de Grenoble (ISTG), University Joseph Fourier of Grenoble, 38031 Grenoble Cedex, France.

K. Nayebi is with the Electrical Engineering Department, Sharif University of Technology, Tehran, Iran.

Digital Object Identifier 10.1109/LSP.2003.819344

Definition 2: The MSF of \mathbf{x} is the vector of score functions of its components, i.e., $\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x}) = (\psi_1(x_1), \dots, \psi_N(x_N))^T$, where

$$\psi_i(x_i) = -\frac{d}{dx_i} \ln p_{x_i}(x_i) = -\frac{p'_{x_i}(x_i)}{p_{x_i}(x_i)} \quad (5)$$

and $p_{x_i}(x_i)$ is the marginal pdf of x_i .

Definition 3: The JSF of the random vector \mathbf{x} is the vector function $\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \dots, \varphi_N(\mathbf{x}))^T$, where

$$\varphi_i(\mathbf{x}) = -\frac{\partial}{\partial x_i} \ln p_{\mathbf{x}}(\mathbf{x}) = -\frac{\frac{\partial}{\partial x_i} p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{x}}(\mathbf{x})} \quad (6)$$

and $p_{\mathbf{x}}(\mathbf{x})$ is the joint pdf of the random vector \mathbf{x} .

Definition 4: The SFD of \mathbf{x} is the difference between its MSF and JSF, i.e., $\boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x}) = \boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x}) - \boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x})$.

The following theorem can be easily proven [5].

Theorem 1: The components of a random vector are independent if, and only if, its SFD is zero.

III. DIFFERENTIAL OF THE MUTUAL INFORMATION

The main theorem of the letter can now be stated.

Theorem 2 (Differential of Mutual Information): Let \mathbf{x} be a bounded random vector, and let $\boldsymbol{\Delta}$ be a “small” random vector with the same dimension. Then

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

where $\boldsymbol{\beta}_{\mathbf{x}}$ is the SFD of \mathbf{x} , and $o(\boldsymbol{\Delta})$ denotes higher order terms in $\boldsymbol{\Delta}$.

Remark 1: Equation (7) may be stated in the following form (which is similar to what is done in [6]):

$$I(\mathbf{x} + \mathcal{E}\mathbf{y}) - I(\mathbf{x}) = E \left\{ (\mathcal{E}\mathbf{y})^T \boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x}) \right\} + o(\mathcal{E}) \quad (8)$$

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 (7) the term “small random vector” is somewhat ad hoc. Conversely, (7) is simpler, and easier to be used in developing gradient-based algorithms for optimizing a mutual information.

Remark 2: Recall that for any multivariate differentiable function $f(\mathbf{x})$, we have

$$f(\mathbf{x} + \boldsymbol{\Delta}) - f(\mathbf{x}) = \boldsymbol{\Delta}^T \nabla f(\mathbf{x}) + o(\boldsymbol{\Delta}). \quad (9)$$

A comparison between (7) and (9) shows that SFD can be called the “stochastic gradient” of the mutual information (although, it must be noted that in (7), \mathbf{x} and $\boldsymbol{\Delta}$ are random vectors, but in (9) they are deterministic vectors).

To prove the theorem, we first have to prove two lemmas. The scalar versions of these lemmas have been already proposed [7].

Lemma 1: Let $\mathbf{x} = (x_1, \dots, x_N)^T$ be a bounded random vector and $\boldsymbol{\Delta} = (\Delta_1, \dots, \Delta_N)^T$ be a “small” random vector. Then

$$p_{\mathbf{x}+\boldsymbol{\Delta}}(\mathbf{t}) - p_{\mathbf{x}}(\mathbf{t}) = -\sum_{i=1}^N \frac{\partial}{\partial t_i} \{E_{\Delta_i} \{\Delta_i | \mathbf{x} = \mathbf{t}\} p_{\mathbf{x}}(\mathbf{t})\} + o(\boldsymbol{\Delta}). \quad (10)$$

Proof: For any differentiable function $h(\mathbf{t})$

$$h(\mathbf{x} + \boldsymbol{\Delta}) - h(\mathbf{x}) = \sum_i \Delta_i \frac{\partial h}{\partial t_i}(\mathbf{x}) + o(\boldsymbol{\Delta}). \quad (11)$$

Thus

$$E \{h(\mathbf{x} + \boldsymbol{\Delta}) - h(\mathbf{x})\} = \sum_i E \left\{ \Delta_i \frac{\partial h}{\partial t_i}(\mathbf{x}) \right\} + o(\boldsymbol{\Delta}). \quad (12)$$

From the well-known [8] relation $E\{g(\mathbf{x}, \mathbf{y})\} = E_{\mathbf{x}}\{E_{\mathbf{y}}\{g(\mathbf{x}, \mathbf{y})|\mathbf{x}\}\}$, the term under summation in the above equation can be written as follows:

$$\begin{aligned} E \left\{ \Delta_i \frac{\partial h}{\partial t_i}(\mathbf{x}) \right\} &= E_{\mathbf{x}} \left\{ E_{\Delta_i} \left\{ \Delta_i \frac{\partial h}{\partial t_i}(\mathbf{x}) | \mathbf{x} \right\} \right\} \\ &= E_{\mathbf{x}} \left\{ \frac{\partial h}{\partial t_i}(\mathbf{x}) E_{\Delta_i} \{\Delta_i | \mathbf{x}\} \right\} \\ &= \int_{\mathbf{t}} \frac{\partial h}{\partial t_i}(\mathbf{t}) E_{\Delta_i} \{\Delta_i | \mathbf{x} = \mathbf{t}\} p_{\mathbf{x}}(\mathbf{t}) d\mathbf{t} \\ &= - \int_{\mathbf{t}} h(\mathbf{t}) \frac{\partial}{\partial t_i} \{E_{\Delta_i} \{\Delta_i | \mathbf{x} = \mathbf{t}\} p_{\mathbf{x}}(\mathbf{t})\} d\mathbf{t}. \end{aligned} \quad (13)$$

The last equality is written by using integration by parts. On the other hand, we have

$$E \{h(\mathbf{x} + \boldsymbol{\Delta}) - h(\mathbf{x})\} = \int_{\mathbf{t}} h(\mathbf{t}) (p_{\mathbf{x}+\boldsymbol{\Delta}}(\mathbf{t}) - p_{\mathbf{x}}(\mathbf{t})) d\mathbf{t}. \quad (14)$$

Now, by combining (12)–(14), we conclude

$$\begin{aligned} &\int_{\mathbf{t}} h(\mathbf{t}) (p_{\mathbf{x}+\boldsymbol{\Delta}}(\mathbf{t}) - p_{\mathbf{x}}(\mathbf{t})) d\mathbf{t} \\ &= - \int_{\mathbf{t}} h(\mathbf{t}) \sum_{i=1}^N \frac{\partial}{\partial t_i} \{E_{\Delta_i} \{\Delta_i | \mathbf{x} = \mathbf{t}\} p_{\mathbf{x}}(\mathbf{t})\} d\mathbf{t} + o(\boldsymbol{\Delta}) \end{aligned} \quad (15)$$

Equation (10) can be deduced from the fact that the above equality holds for any function h . ■

Lemma 2: Let \mathbf{x} and $\boldsymbol{\Delta}$ be as defined in Lemma 1. Then

$$H(\mathbf{x} + \boldsymbol{\Delta}) - H(\mathbf{x}) = -E \left\{ \boldsymbol{\Delta}^T \boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x}) \right\} + o(\boldsymbol{\Delta}) \quad (16)$$

where H denotes Shannon’s entropy, and $p_{\mathbf{x}}(\cdot)$ and $\boldsymbol{\varphi}_{\mathbf{x}}(\cdot)$ are the pdf and the JSF of \mathbf{x} , respectively.

Proof: We write

$$\begin{aligned} H(\mathbf{x} + \boldsymbol{\Delta}) - H(\mathbf{x}) &= -E \{ \ln p_{\mathbf{x}+\boldsymbol{\Delta}}(\mathbf{x} + \boldsymbol{\Delta}) \} + E \{ \ln p_{\mathbf{x}}(\mathbf{x}) \} \\ &= E \left\{ \ln \frac{p_{\mathbf{x}}(\mathbf{x} + \boldsymbol{\Delta})}{p_{\mathbf{x}+\boldsymbol{\Delta}}(\mathbf{x} + \boldsymbol{\Delta})} \right\} \\ &\quad - E \left\{ \ln \frac{p_{\mathbf{x}}(\mathbf{x} + \boldsymbol{\Delta})}{p_{\mathbf{x}}(\mathbf{x})} \right\}. \end{aligned} \quad (17)$$

In the neighborhood of 1, $\ln x = (x-1) - (1/2)(x-1)^2 + \dots$, and hence by defining $\mathbf{z} \triangleq \mathbf{x} + \Delta$, the first term of (17) can be written as

$$\begin{aligned} E \left\{ \ln \frac{p_{\mathbf{x}}(\mathbf{z})}{p_{\mathbf{z}}(\mathbf{z})} \right\} &= E \left\{ \frac{p_{\mathbf{x}}(\mathbf{z})}{p_{\mathbf{z}}(\mathbf{z})} - 1 \right\} + o(\Delta) \\ &= \int_{\mathbf{t}} \left(\frac{p_{\mathbf{x}}(\mathbf{t})}{p_{\mathbf{z}}(\mathbf{t})} - 1 \right) p_{\mathbf{z}}(\mathbf{t}) d\mathbf{t} + o(\Delta) \\ &= o(\Delta). \end{aligned} \quad (18)$$

The second right term of (17) is simplified as follows:

$$\begin{aligned} &- E \left\{ \ln \frac{p_{\mathbf{x}}(\mathbf{x} + \Delta)}{p_{\mathbf{x}}(\mathbf{x})} \right\} \\ &= E \{ \ln p_{\mathbf{x}}(\mathbf{x}) \} - E \{ \ln p_{\mathbf{x}}(\mathbf{x} + \Delta) \} \\ &= \int_{\mathbf{t}} \ln p_{\mathbf{x}}(\mathbf{t}) p_{\mathbf{x}}(\mathbf{t}) d\mathbf{t} - \int_{\mathbf{t}} \ln p_{\mathbf{x}}(\mathbf{t}) p_{\mathbf{x}+\Delta}(\mathbf{t}) d\mathbf{t} \\ &= \int_{\mathbf{t}} \ln p_{\mathbf{x}}(\mathbf{t}) (p_{\mathbf{x}}(\mathbf{t}) - p_{\mathbf{x}+\Delta}(\mathbf{t})) d\mathbf{t} \\ &= \sum_i \int_{\mathbf{t}} \ln p_{\mathbf{x}}(\mathbf{t}) \frac{\partial}{\partial t_i} \{ E_{\Delta_i} \{ \Delta_i | \mathbf{x} = \mathbf{t} \} p_{\mathbf{x}}(\mathbf{t}) \} d\mathbf{t} + o(\Delta) \\ &\quad (\text{using Lemma 1}) \\ &= - \sum_i \int_{\mathbf{t}} E_{\Delta_i} \{ \Delta_i | \mathbf{x} = \mathbf{t} \} \varphi_i(\mathbf{t}) p_{\mathbf{x}}(\mathbf{t}) d\mathbf{t} + o(\Delta) \\ &\quad (\text{integration by parts}) \\ &= - \sum_i E_{\mathbf{x}} \{ E_{\Delta_i} \{ \Delta_i | \mathbf{x} \} \varphi_i(\mathbf{x}) \} + o(\Delta) \\ &= - \sum_i E \{ \Delta_i \varphi_i(\mathbf{x}) \} + o(\Delta) \\ &= - E \left\{ \Delta^T \boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x}) \right\} + o(\Delta). \end{aligned} \quad (19)$$

The lemma is proven by combining (17)–(19). ■

Corollary 1: For scalar random variables x_i and Δ_i , we have

$$H(x_i + \Delta_i) - H(x_i) = -E \{ \Delta_i \cdot \psi_{x_i}(x_i) \} + o(\Delta_i). \quad (20)$$

Proof of theorem 2: Combining the usual expression $I(\mathbf{x}) = \sum_i H(x_i) - H(\mathbf{x})$ with (16) and (20) proves the theorem. ■

IV. APPLICATION IN BSS

In this section, we use the results of the previous section for deriving estimation equations for source separation in linear mixtures.

A. Linear Instantaneous Mixtures

We first calculate the gradient of $I(\mathbf{y})$ with respect to the separating matrix \mathbf{B} . Let $\hat{\mathbf{B}} = \mathbf{B} + \mathcal{E}$, where $\mathcal{E} = [\epsilon_{ij}]$ is a “small” matrix. The new output vector is $\hat{\mathbf{y}} = \mathbf{y} + \mathcal{E}\mathbf{x}$. From Theorem 2, the variation of I will be (up to first-order terms)

$$I(\hat{\mathbf{y}}) - I(\mathbf{y}) = E \left\{ \boldsymbol{\beta}_{\mathbf{y}}^T(\mathbf{y}) \mathcal{E} \mathbf{x} \right\} = \sum_{i,j} \epsilon_{ij} E \{ \beta_{\mathbf{y},i}(\mathbf{y}) x_j \} \quad (21)$$

where $\boldsymbol{\beta}_{\mathbf{y}}$ denotes the SFD of \mathbf{y} . This equation shows that $(\partial I / \partial b_{ij}) = E \{ \beta_{\mathbf{y},i}(\mathbf{y}) x_j \}$, and hence

$$\frac{\partial I}{\partial \mathbf{B}} = E \{ \boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y}) \mathbf{x}^T \}. \quad (22)$$

Finally, the steepest descent algorithm for estimating the matrix \mathbf{B} is

$$\mathbf{B} \leftarrow \mathbf{B} - \mu E \{ \boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y}) \mathbf{x}^T \}. \quad (23)$$

B. Linear Convolutional Mixtures

In this section, we show how the theorem can be used in separating convolutional mixtures, i.e., when the mixing matrix is composed of linear-time-invariant filters instead of scalars. Suppose that the separating filters are FIR with maximum degree M . Then, the separating matrix is in the form $\mathbf{B}(z) = \sum_{k=0}^M \mathbf{B}_k z^{-k}$, and the output vector is

$$\mathbf{y}(n) = \mathbf{B}_0 \mathbf{x}(n) + \mathbf{B}_1 \mathbf{x}(n-1) + \dots + \mathbf{B}_M \mathbf{x}(n-M). \quad (24)$$

For separating the sources, $\mathbf{B}_0, \dots, \mathbf{B}_M$ must be determined to produce independent outputs. Here, a simple multiplicative relation like (2) does not exist, and the traditional method fails in calculating the gradient of $I(\mathbf{y}(n))$ with respect to \mathbf{B}_k . But, by using Theorem 2, this gradient can be easily computed. First, we write $\hat{\mathbf{B}}_k = \mathbf{B}_k + \mathcal{E}$ and then

$$I(\hat{\mathbf{y}}(n)) - I(\mathbf{y}(n)) = E \left\{ \boldsymbol{\beta}_{\mathbf{y}}^T(\mathbf{y}(n)) \mathcal{E} \mathbf{x}(n-k) \right\} \quad (25)$$

and from there

$$\frac{\partial I(\mathbf{y}(n))}{\partial \mathbf{B}_k} = E \left\{ \boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y}(n)) \mathbf{x}(n-k)^T \right\}. \quad (26)$$

However, in convolutional mixtures, instantaneous independence is not sufficient for separating the sources, and using the above gradient needs some more considerations, detailed for instance in [5]. As shown in [5], the whole criterion requires a few terms that can be computed with equations similar to (26), and it leads to an efficient algorithm.

V. EXPERIMENTAL RESULTS

Here, we present separation results for linear instantaneous mixtures. Sources are a sine wave and a uniform random signal, both with zero mean and unit variance. The mixing matrix is

$$\mathbf{A} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}.$$

For using the algorithm (23), we need to estimate the SFD of \mathbf{y} . For this estimation, we have used a simple histogram estimation method (but other estimators could be used). In this method, y_1 and y_2 are first splitted into some bins. Let N and $\text{Card}(n_1, n_2)$ denote, respectively, the total number of output samples and the number of samples in the bin (n_1, n_2) . Then $p(n_1, n_2) = (\text{Card}(n_1, n_2)/N)$ is the joint probability estimation in (n_1, n_2) , and $p(n_2|n_1) = (p(n_1, n_2)/\sum_{n_2} p(n_1, n_2))$ is the conditional probability estimation. Finally, noting that $\beta_1(\mathbf{y}) = (\partial/\partial y_1) \ln p(y_2|y_1) = (\partial/\partial y_1) p(y_2|y_1)/p(y_2|y_1)$, we can estimate $\beta_1(n_1)$ as

$$\hat{\beta}_1(n_1) = \frac{p(n_2|n_1) - p(n_2|n_1 - 1)}{p(n_2|n_1)}. \quad (27)$$

$\beta_2(n_2)$ will be estimated in a similar way.

We used $N = 500$ and $\mu = 0.1$. For estimating the SFD, y_1 and y_2 are split into ten bins each (i.e., a 10×10 histogram is used). The initial value of \mathbf{B} is the identity matrix, and the expectation in (23) is estimated by the empirical average on the data block.

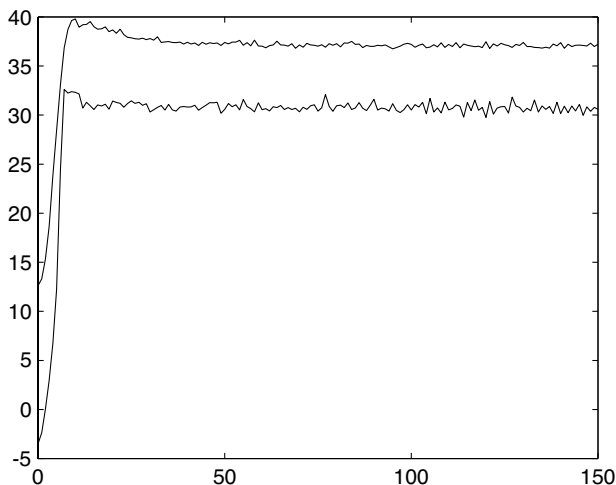


Fig. 1. Output SNRs in separating the mixture of two sources.

For measuring the quality of separation, the output SNR is used, which is defined by (in decibels)

$$\text{SNR} = 10 \log_{10} \frac{E\{s^2\}}{E\{(s-y)^2\}}. \quad (28)$$

Fig. 1 shows the averaged output SNRs versus iteration, taken over 100 runs of the algorithm. As can be seen in the figure, a good separation quality is obtained: 37 and 31 dB. If the same experiment is repeated using the method of [3], with a ten-bin histogram estimation for marginal score functions, the averaged output SNRs will be 29 and 17 dB. Moreover, if in the method of [3], the optimal estimation of marginal score functions using a third-order polynomial is used, then the averaged output SNRs will be 44 and 35 dB. It is clear, then, that this new approach based on SFD is much more efficient than approaches based only on marginal score functions, in the sense that it performs much better despite crude density (and score function) estimations. Finally, it must be emphasized that the main advantage of the new method is its generality: it can be easily extended to more general mixing models, e.g., convolutive mixtures [5] and convolutive post-nonlinear mixtures [9]

The good quality of the approach based on SFD can be explained as follows. First, note that in [3], the authors use the natural gradient [10]

$$\begin{aligned} \nabla_{\mathbf{B}} I &= \frac{\partial I}{\partial \mathbf{B}} \cdot \mathbf{B}^T \\ &= E\{\boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\} \\ &= E\{\boldsymbol{\psi}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\} - E\{\boldsymbol{\varphi}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\}. \end{aligned} \quad (29)$$

However, integration by parts shows that $E\{\boldsymbol{\varphi}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\} = \mathbf{I}$, and hence

$$\nabla_{\mathbf{B}} I = E\{\boldsymbol{\psi}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\} - \mathbf{I}. \quad (30)$$

Practically, this equation is simpler than (29), since it only requires estimation of marginal score functions. However, the separation information is contained in the averaged SFD, as the gradient of the mutual information. Moreover, SFD is the difference of two terms $\boldsymbol{\psi}_{\mathbf{y}}(\mathbf{y})$ and $\boldsymbol{\varphi}_{\mathbf{y}}(\mathbf{y})$, and consequently $\nabla_{\mathbf{B}} I$ will be the difference of the two terms $E\{\boldsymbol{\psi}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\}$ and $E\{\boldsymbol{\varphi}_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\}$. In (30), one of these terms is exactly computed. Therefore, a good estimation of the other term is required for separating the sources, because the difference of these terms must vanish for achieving the convergence in a gradient-based algorithm (note also that from Theorem 1, the separation achieves when the SFD of outputs vanishes). However, in the method presented in this letter, we directly estimate the SFD, and hence a good separation can be achieved even with a simple histogram approximation.

VI. CONCLUSION

In this letter, the variation of the mutual information resulting from a small variation in its argument (the “differential” of the mutual information) has been calculated. It can be used for developing gradient-based algorithms in any mutual information optimization problem. As an example, we used it for developing a new algorithm for blind source separation. Experimental results, for linear instantaneous mixtures, show the good performance of the resulting algorithm. The application of the method for separating more general mixtures is currently under study.

REFERENCES

- [1] P. Comon, “Independent component analysis, a new concept?,” *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.
- [2] D. T. Pham, “Blind separation of instantaneous mixtures via an independent component analysis,” *IEEE Trans. Signal Processing*, vol. 44, pp. 2768–2779, Nov. 1996.
- [3] A. Taleb and C. Jutten, “Entropy optimization, application to blind source separation,” in *Proc. ICANN*, Lausanne, Switzerland, Oct. 1997, pp. 529–534.
- [4] A. Taleb and C. Jutten, “Source separation in post nonlinear mixtures,” *IEEE Trans. Signal Processing*, vol. 47, pp. 2807–2820, Oct. 1999.
- [5] M. Babaie-Zadeh, C. Jutten, and K. Nayebi, “Separating convolutive mixtures by mutual information minimization,” in *Proc. IWANN*, Granada, Spain, June 2001, pp. 834–842.
- [6] D. T. Pham, “Mutual information approach to blind separation of stationary sources,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 1–12, July 2002.
- [7] S. Achard, *Initiation à la séparation aveugle de sources dans des mélanges post non linéaires*, France: DEA de Insti. Nat. Polytechnique de Grenoble, June 2000.
- [8] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1991.
- [9] M. Babaie-Zadeh, C. Jutten, and K. Nayebi, “Blind separating convolutive post-nonlinear mixtures,” in *Proc. ICA*, San Diego, CA, Dec. 2001, pp. 138–143.
- [10] J.-F. Cardoso and B. Laheld, “Equivariant adaptive source separation,” *IEEE Trans. Signal Processing*, vol. 44, pp. 3017–3030, Dec. 1996.