A NEW METHOD FOR ESTIMATING SCORE FUNCTION DIFFERENCE (SFD) AND ITS APPLICATION TO BLIND SOURCE SEPARATION

Bahman Bahmani[1]*, Massoud Babaie-Zadeh*[1]*, and Christian Jutten*[2]

[1] Advanced Communication Research Institute (ACRI), Electrical Engineering Department,
Sharif University of Technology, Tehran, Iran.
[2] Laboratoire des Images et des Signaux (LIS), Institut National Polytechnique
de Grenoble (INPG), Grenoble, France.
bahmanibahman@yahoo.com , mbzadeh@yahoo.com, Christian.Jutten@inpg.fr

ABSTRACT

Score Function Difference (SFD) is a recently proposed "gradient" for mutual information which can be used in Blind Source Separation algorithms based on minimization of mutual information. To be applied to practical problems, SFD must be estimated from the data samples. In this paper, a new method for estimating SFD is proposed. To compare the performance of this new estimator with other proposed SFD estimation methods, we have applied them in separating linear instantaneous mixtures. It will be seen that our method performs superior to all other methods previously proposed for estimation of SFD.

1. INTRODUCTION

Blind Source Separation (BSS) [1, 2] 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. Since the only information about source signals is their statistical independence, a general approach for BSS is to design the separating system which transforms again the observations to statistically independent outputs. This approach is called Independent Component Analysis (ICA), and for linear mixtures, it is shown to result in retrieving the sources up to some trivial indeterminacies [3].

ICA can be obtained by optimizing a "contrast function" *i.e.* a scalar measure of the independence of the outputs [4, 3]. One of the widely used contrast functions is mutual information, which has been shown [4] to provide an asymptotically Maximum-Likelihood (ML) estimation of source signals in linear instantaneous mixtures. Recently, a non-parametric "gradient" for mutual information, called Score Function Difference (SFD), has been proposed [5]. SFD has been used successfully in separating different mixing models [6]. To be applied to practical problems, the algorithms based on SFD need its data derived estimation.

In this paper, a new method for estimating SFD is proposed and it will be applied to blind source separation of linear instantaneous mixtures. This new method is shown to perform superior to all previously proposed SFD estimation methods.

The paper is organized as follows. Section 2 reviews the essential materials to express the "gradient" of mutual information. The new method for estimating SFD is developed in Section 3. Section 4 presents some experimental results. Finally, conclusions are made in Section 5.

This work has been partially funded by Sharif University of Technology, by French Embassy in Tehran, and by Center for International Research and Collaboration (ISMO).

2. PRELIMINARY ISSUES

2.1 Mutual information

For designing a system which generates independent outputs, we need a criterion for measuring their independence. Recall that random variables $y_1,\dots,y_N$ are independent if and only if $p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{N} p_{y_i}(y_i)$, where p stands for the Probability Density Function (PDF) . A convenient independence measure is mutual information [7] of y_i's, denoted by $I(\mathbf{y})$, which is the Kullback-Leibler divergence between $p_{\mathbf{y}}(\mathbf{y})$ and $\prod_{i=1}^{N} p_{y_i}(y_i)$:

$$\begin{aligned} I(\mathbf{y}) &= D(p_{\mathbf{y}}(\mathbf{y}) \,\|\, \prod_{i=1}^{N} p_{y_i}(y_i)) \\ &= \int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) \ln \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_{i=1}^{N} p_{y_i}(y_i)}\, d\mathbf{y} \end{aligned} \tag{1}$$

It is well-known that this quantity is always non-negative, and vanishes if and only if the y_i's are independent. Consequently, the parameters of the separating system can be calculated based on minimization of the mutual information of the outputs.

To do this minimization, knowing an expression for the "gradient" of the mutual information is helpful. Such an expression, which has been already proposed [5], requires multivariate score functions.

2.2 "Gradient" of mutual information

The variations of mutual information resulted from a small deviation in its argument (the "differential" of mutual information), is given by the following theorem [5]:

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

$$I(\mathbf{y}+\Delta) - I(\mathbf{y}) = E\left\{\Delta^T \boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y})\right\} + o(\Delta) \tag{2}$$

where $o(\Delta)$ denotes higher order terms in Δ.

In this Theorem, the function $\boldsymbol{\beta}_{\mathbf{y}}(\mathbf{y})$, called Score Function Difference (SFD) [8], is defined as follows.

Definition 1 (SFD) *The score function difference (SFD) of a random vector* $\mathbf{y}$ *is the difference between its marginal score function* $\psi_{\mathbf{y}}(\mathbf{y})$ *(MSF) and joint score function* $\varphi_{\mathbf{y}}(\mathbf{y})$ *(JSF):*

$$\beta_{\mathbf{y}}(\mathbf{y}) = \psi_{\mathbf{y}}(\mathbf{y}) - \varphi_{\mathbf{y}}(\mathbf{y}) \tag{3}$$

where the marginal score function is defined by

$$\psi_{\mathbf{y}}(\mathbf{y}) = \big(\psi_1(y_1), \ldots, \psi_N(y_N)\big)^T \tag{4}$$

with

$$\psi_i(y_i) = -\frac{d}{dy_i} \ln p_{y_i}(y_i) = -\frac{p'_{y_i}(y_i)}{p_{y_i}(y_i)}. \tag{5}$$

and the joint score function is defined by

$$\varphi_{\mathbf{y}}(\mathbf{y}) = \big(\varphi_1(\mathbf{y}), \ldots, \varphi_N(\mathbf{y})\big)^T \tag{6}$$

with

$$\varphi_i(\mathbf{y}) = -\frac{\partial}{\partial y_i} \ln p_{\mathbf{y}}(\mathbf{y}) = -\frac{\frac{\partial}{\partial y_i} p_{\mathbf{y}}(\mathbf{y})}{p_{\mathbf{y}}(\mathbf{y})} \tag{7}$$

SFD plays an important role for minimizing the mutual information. In fact, for any multivariate differentiable function $f(\mathbf{y})$, we have:

$$f(\mathbf{y}+\Delta) - f(\mathbf{y}) = \Delta^T \nabla f(\mathbf{y}) + o(\Delta) \tag{8}$$

Then, a comparison between (2) and (8) shows that the so-called SFD can be called the *stochastic gradient* of the mutual information.

2.3 Relation between SFD and conditional densities

One of the many properties of SFD, which has been proved in [9], is that it can be stated in terms of conditional densities as follows:

Property 1 *For a random vector* $\mathbf{y} = (y_1, \cdots, y_N)^T$ *we have:*

$$\begin{aligned} \beta_i(\mathbf{y}) &= \frac{\partial}{\partial y_i} \big\{ \ln p(y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N | y_i) \big\} \\ &= \frac{\frac{\partial}{\partial y_i} p(y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N | y_i)}{p(y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N | y_i)} \end{aligned} \tag{9}$$

where $\beta_i(\mathbf{y})$ *denotes the i-th component of the SFD of* $\mathbf{y}$.

As we will see in the next Section, this property, though very simple, proves to do a good job in estimating SFD.

3. PROPOSED SFD ESTIMATION METHOD

Our main idea for estimating SFD is to use Property 1. Contrary to the definition of SFD (definition 1), which gives SFD indirectly from the functions MSF and JSF, this property gives a representation for SFD which makes it possible to estimate SFD directly (not from estimations of MSF and JSF, as it is done in all previously proposed methods for estimating SFD). Actually, if we first independently estimate MSF and JSF, there will be estimation errors in both estimations, which results in poorer estimation of SFD [9], but these "double" errors do not exist in dependent estimations (MSF can be estimated by integration of the JSF estimate) [9, 10] or in the direct estimation we propose based on Property 1.

3.1 Estimating conditional densities and their derivatives

It is seen from (9) that if we estimate conditional densities of the form $p(\mathbf{y}|x)$, where $\mathbf{y}$ is a random vector and x is a random variable, and their derivative with respect to the conditioning variable (x), then we will be able to make an estimation of SFD. In statistics literature, there are various methods for estimating conditional densities. However, we need not only the estimation of conditional densities, but also the estimation of their derivatives. Among the existing methods for conditional density estimation, the method proposed in [11] can be easily adjusted to be applied to our problem. The resulting method is as follows.

Let $\mathbf{y}$ be a d-dimensional random vector, x be a random variable and $p(\mathbf{y}|x)$ be the conditional density of $\mathbf{y}$ given x, which is assumed smooth in both x and $\mathbf{y}$. Furthermore, let $p^{(i)}(\mathbf{y}|x)$ denote the i-th derivative of $p(\mathbf{y}|x)$ with respect to x. The goal is to estimate functions $p(\mathbf{y}|x)$ and $p^{(1)}(\mathbf{y}|x)$ based on a sequence of observations $(\mathbf{y}_1, x_1), \ldots, (\mathbf{y}_n, x_n)$.

As it is mentioned in [11], estimating the conditional density (and its derivatives) can be regarded as a non-parametric regression problem. To make this connection, note that:

$$E\left\{\delta(\mathbf{y}-\mathbf{y}_0)|x=x_0\right\} = \int_{\mathbf{y}} \delta(\mathbf{y}-\mathbf{y}_0)\, p(\mathbf{y}|x_0)\, d\mathbf{y} = p(\mathbf{y}_0|x_0) \tag{10}$$

where $\delta(\mathbf{y})$ is the Dirac delta function. From this equation it can be deduced that:

$$E\left\{K_b(\mathbf{y}-\mathbf{y}_0)|x=x_0\right\} \simeq p(\mathbf{y}_0|x_0) \quad \text{as} \quad b \to 0 \tag{11}$$

where K is a 'kernel' function, that is, the density of a d-dimensional random vector with zero mean and unit variance, and $K_b(\mathbf{u}) = b^{-d} \cdot K(\frac{\mathbf{u}}{b})$ (b is usually called the 'bandwidth'). Now, by using the r-th order Taylor's expansion about x_0, we have:

$$\begin{aligned} E\left\{K_b(\mathbf{y}-\mathbf{y}_0)|x=z\right\} &\simeq p(\mathbf{y}_0|z) \\ &\simeq \sum_{i=0}^{r} \left\{ \frac{1}{i!} p^{(i)}(\mathbf{y}_0|x_0)(z-x_0)^i \right\} \end{aligned} \tag{12}$$

This suggests the following least squares problem. Let $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_r)$ minimize:

$$R(\theta; x_0, \mathbf{y}_0) = \sum_{i=1}^{n} \left\{ K_b(\mathbf{y}_i - \mathbf{y}_0) - \sum_{j=0}^{r} \theta_j (x_i - x_0)^j \right\}^2 \cdot W_h(x_i - x_0) \tag{13}$$

Where W is a symmetric scalar density function and $W_h(x) = h^{-1}.W(\frac{x}{h})$. Then recalling equation (12) we can estimate $p(\mathbf{y}_0|x_0) \simeq \hat{\theta}_0$ and $p^{(1)}(\mathbf{y}_0|x_0) \simeq \hat{\theta}_1$.

3.2 Estimating SFD

Now, having an estimation for conditional density and its derivative (with respect to the conditioning variable), we return to the main problem of estimating SFD.

Let $\mathbf{y} = (y_1, \cdots, y_N)^T$ be a random vector and define the notation $\mathbf{y}^{(i)} \triangleq (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N)^T$, for $i = 1, \cdots, N$. Furthermore, assume that we have at hand n samples of $\mathbf{y}$, namely $\mathbf{y}(1), \cdots, \mathbf{y}(n)$. Thus, for each $1 \le i \le n$ we have n

samples of $\mathbf{y}^{(i)}$ (that is, $\mathbf{y}^{(i)}(1),\cdots,\mathbf{y}^{(i)}(n)$) and n samples of y_i (that is, $y_i(1),\cdots,y_i(n)$). Now to estimate the value of $\beta_i(\mathbf{y})$ $(1 \le i \le n)$, we note that from Property 1 we have:

$$\beta_i(\mathbf{y}) = \frac{\frac{\partial}{\partial y_i} p(\mathbf{y}^{(i)}|y_i)}{p(\mathbf{y}^{(i)}|y_i)} \tag{14}$$

Now, using n sample pairs $(\mathbf{y}^{(i)}(1),y_i(1)),\cdots,(\mathbf{y}^{(i)}(n),y_i(n))$ we can estimate each of the numerator and the denominator of the fraction in equation (14) by the explained method. This results in estimation of SFD.

4. EXPERIMENTAL RESULTS

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

$$\mathbf{A} = \begin{bmatrix} 1 & 0.7 \\ 0.5 & 1 \end{bmatrix} \tag{15}$$

For separating the sources we used the SFD-based algorithm proposed in [6]. This algorithm is briefly as follows:

- Initialization: $\mathbf{B} = \mathbf{I}$.
- Loop:
  1. $\mathbf{y} = \mathbf{B}\mathbf{x}$.
  2. Estimate $\beta_{\mathbf{y}}(\mathbf{y})$.
  3. $\nabla_{\mathbf{B}} I = \widehat{E}\left\{\beta_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T\right\}$
  4. $\mathbf{B} \leftarrow (\mathbf{I} - \mu\nabla_{\mathbf{B}} I)\,\mathbf{B}$.
  5. Normalization: Divide the i-th row of the matrix $\mathbf{B}$ by σ_i, where σ_i^2 is the energy of y_i.
- Repeat until convergence.

In the above algorithm, $\mathbf{B}$ is the separating matrix, $\mathbf{I}$ denotes the identity matrix, $\mathbf{x}$ stands for the observation vector, $\mathbf{y}$ denotes the output vector, and I is the mutual information of the outputs.

Three methods for estimating SFD, a kernel based estimator, a histogram based estimator, and a polynomial estimator, are proposed in [6] and another method for this estimation is proposed by D.-T. Pham in [12]. We implemented the above algorithm with these four SFD estimation methods and with the estimation method proposed in this paper. For all of these implementations $\mu = 0.1$ was chosen. Besides, we took $r = 2$ and $b = h = 0.3$, with Gaussian densities as kernel function ($K(.)$) and weight function ($W(.)$), for the proposed method. Figures 1 through 5 show the averaged Signal To Noise Ratios (SNR's), taken over 100 experiments, versus iteration for these methods. SNR is defined as:

$$\text{SNR} = \frac{\text{SNR}_1 + \text{SNR}_2}{2} \tag{16}$$

where (assuming no permutation):

$$\text{SNR}_i(\text{in dB}) = 10\log_{10}\frac{E\left\{s_i^2\right\}}{E\left\{(y_i - s_i)^2\right\}}, \qquad i = 1,2 \tag{17}$$

Furthermore, Table 1 shows, for each estimation method, the average and the variance of the SNR's (after convergence), taken over 100 runs of the algorithm. As it is seen in this table, our method has better separation performance than all other methods previously developed for SFD estimation. Moreover, as it can be seen from Figures 1 to 5, the separation algorithm when using our method converges in much fewer iterations than when other methods are applied.

However, our method is computationally more demanding. To have an idea about the computational load of these methods, we have listed in Table 2 the average time needed for convergence of each method in our implementation (which was done using MATLAB6.1 on a 802MHz Pentium III machine with 256MB RAM and Windows XP platform). Although, the comparison based on this table is not very accurate, it roughly shows that the proposed method has more computational load than other methods.

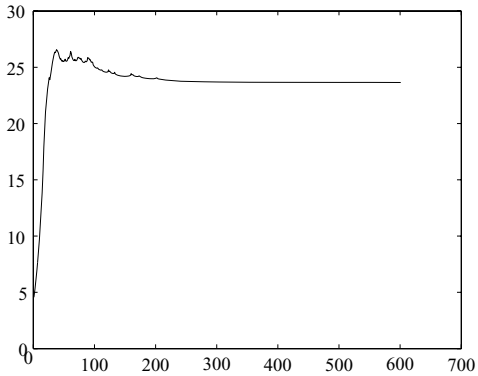

Figure 1: Kernel method.

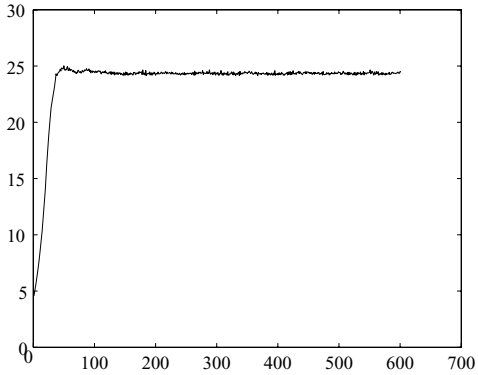

Figure 2: Histogram method.

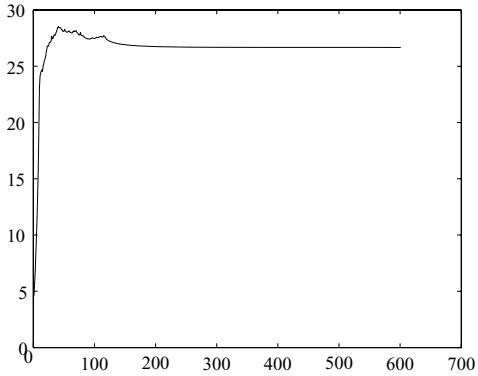

Figure 3: Pham's method.

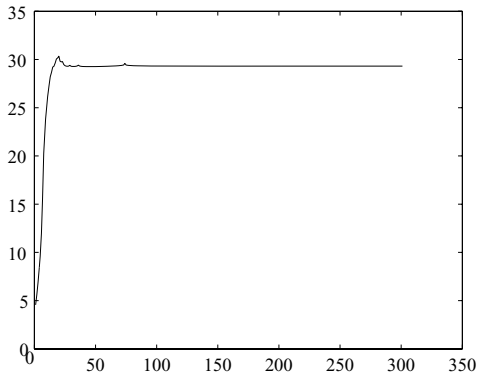

Figure 4: Polynomial method.

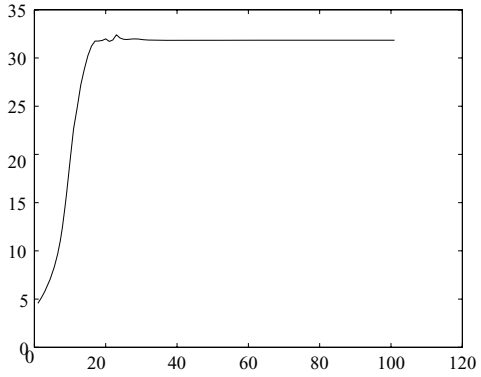

Figure 5: Our method.

5. CONCLUSION

In this paper a new method for estimating Score Function Difference (SFD) has been proposed. The method is based on a direct representation of SFD in terms of conditional densities and an adjusted conditional density estimation method. The proposed estimation method has been applied to blind separation of linear instantaneous mixtures. It has been shown that the proposed method performs superior to all previously developed SFD estimation methods. Furthermore, it has been shown that our algorithm needs much fewer iterations for convergence than all other methods.

However, this better performance has been obtained at the expense of a higher computational load. Moreover, the proposed method has two bandwidth parameters (b and h) which should be suitably selected. We selected these parameters in our simulations, mainly by means of trial and error, but we observed that the proposed method is not very dependent on the values of these parameters. Finding a more sophisticated selection of these parameters is currently under study.

Method	Averaged SNR (in dB)	SNR Variance
Kernel method	23.6514	7.8978
Histogram method	24.4580	6.7986
Pham's method	26.6697	8.0193
Polynomial method	29.3120	6.9094
Our method	31.8385	7.6698

Table 1: Averaged (over 100 experiments) output SNR's for different estimation methods for SFD.

Method	Average convergence time(Sec)
Kernel method	3.9287
Histogram method	1.9819
Pham's method	3.8355
Polynomial method	1.3450
Our method	6.8609

Table 2: Average time needed for convergence for different SFD estimation methods.

REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiely & Sons, 2001.

[2] Andrzej Cichocki and Shun-ichi Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley and sons, 2002.

[3] P. Comon, "Independent component analysis, a new concept?", *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[4] J.-F. Cardoso, "Blind signal separation: statistical principles", *Proceedings 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 and C. Jutten, "A general approach for mutual information minimization and its application to blind source separation", To appear in *Signal Processing*.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, 1991.

[8] M. Babaie-Zadeh, C. Jutten, and K. Nayebi, "Blind separating Convolutive Post-Nonlinear mixtures", in *Proceedings of ICA2001*, San Diego (Ca, USA), December 2001, pp. 138–143.

[9] M. Babaie-Zadeh, *On blind source separation in convolutive and nonlinear mixtures*, PhD thesis, INP Grenoble, 2002.

[10] S. Achard, D. T. Pham, and C. Jutten, "Criteria for blind source separation in post nonlinear mixtures with mutual information", To appear in *Signal Processing*.

[11] J. Fan, Q. Yao, and H. Tong, "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems", *Biometrika*, vol. 83, pp. 189–206, 1996.

[12] D. T. Pham, "Fast algorithm for estimating mutual information, entropies and score functions", in *Proceedings of ICA2003*, Nara, Japan, April 2003, pp. 17–22.