IST2005
INTERNATIONAL SYMPOSIUM ON TELECOMMUNICATIONS
September 10-12, 2005 Shiraz, Iran

Performance Comparison Of Different Score Function Difference Estimation Methods

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.
Emails: 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 (BSS) algorithms based on minimization of mutual information. To be applied to practical problems, SFD must be estimated from data samples, and till now there are several algorithms for the estimation of SFD from data. However, the comparison of the performances of these algorithms has never been addressed in the literature. The criterion usually used in the literature for comparing different SFD estimators is the quality of separation in BSS algorithms. But this practical criterion does not show that the considered estimation method makes a really good estimation of the actual SFD. In this paper, a simple method for evaluating the performance of an SFD estimator is proposed, and using it the performances of the currently known SFD estimators are compared.

Keywords: Blind Source Separation, BSS, ICA, Independent Components Analysis, Mutual Information, Score Function Difference estimation, SFD estimation.

*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).

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 (MI), 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 and complicated mixing models [6].

To be applied to practical problems, algorithms based on SFD require its estimation from data. Several methods for estimating SFD have been introduced in the literature [6],[7],[8]. These SFD estimation methods are all applied to the blind source separation problem, and they result in different performances in BSS. However, a comparison of the performance of these different estimators

have never been addressed in the literature.

It should also be noted that when a SFD estimator is used for BSS, the quality of source separation by no means shows that the estimation method is a really good estimator of the actual SFD. For instance, the authors have recently shown [9] that when applied to BSS, a "poor" estimation of SFD may have advantages to a "better" SFD estimator.

One problem in comparing the performances of SFD estimators is that SFD depends on a multivariate probability density function (PDF), and hence its theoretical value (to be served as a basis for comparing the performances of the estimators) is not easy to calculate. In this paper, we are going to propose a simple method for comparing the performances of different SFD estimators, and use it to compare the performances of the currently known SFD estimation algorithms.

The paper is organized as follows. Section 1 reviews the essential materials to express the "gradient" of mutual information. The comparison method is developed in Section 2. Section 3 presents some experimental results. Finally, conclusions are made in Section 4.

1 Preliminary Issues

1.1 Mutual information

For designing a system which generates independent outputs, we need a criterion for measuring their independence. Recall that random variables $x_1, \ldots, x_N$ are independent if and only if $p_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{N} p_{x_i}(x_i)$, where p stands for the Probability Density Function (PDF) . A convenient independence measure is mutual information [10] of x_i's, denoted by $I(\mathbf{x})$, which is the Kullback-Leibler divergence between $p_{\mathbf{x}}(\mathbf{x})$ and $\prod_{i=1}^{N} p_{x_i}(x_i)$:

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

It is well-known that this quantity is always non-negative, and vanishes if and only if the x_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.

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

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

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

In this Theorem, the function $\boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x})$, called Score Function Difference (SFD) [11], is defined as follows.

Definition 1 (SFD) *The score function difference (SFD) of a random vector $\mathbf{x}$ is the difference between its marginal score function $\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x})$ (MSF) and joint score function $\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x})$ (JSF):*

$$\boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x}) = \boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x}) - \boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x}) \tag{3}$$

where the marginal score function is defined by

$$\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x}) = (\psi_1(x_1), \ldots, \psi_N(x_N))^T \tag{4}$$

with

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

and the joint score function is defined by

$$\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_N(\mathbf{x}))^T \tag{6}$$

with

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

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

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

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

2 Proposed comparison method

In this section, we explain our comparison method. First we recall that the SFD of a random vector is

identically zero if and only if its components are statistically independent [6].

Consider now two independent random variables s_1 and s_2 which both have "Bimodal Gaussian" distributions having modes with means $+1$ and -1 and equal variances. In other words:

$$p_{s_i}(s) = \frac{G(s;-1,\sigma^2) + G(s;+1,\sigma^2)}{2}, \quad i = 1,2 \tag{9}$$

where $p_{s_i}(\cdot)$ is the PDF of s_i, and $G(\cdot;\mu,\sigma)$ stands for the PDF of a Gaussian random variable with mean μ and variance σ^2:

$$G(t;\mu,\sigma^2) \triangleq \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} \tag{10}$$

Signals with the PDF of (9) are very common in digital telecommunications: they arise in transmitting one bit of data through an Additive White Gaussian Noise (AWGN) channel.

Because of the independence of s_1 and s_2, the SFD of the random vector $(s_1,s_2)^T$ is equal to zero. Now, we mix the signals s_1 and s_2 by a rotation transformation to obtain statistically dependent random variables x_1 and x_2, i.e.:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \cdot \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \tag{11}$$

where θ is the angle of rotation.

In the following we will theoretically calculate $\boldsymbol{\beta}_{\mathbf{x}}(x_1,x_2) = (\beta_1(x_1,x_2),\beta_2(x_1,x_2))^T$, the SFD of the random vector $\mathbf{x} = (x_1,x_2)^T$. It is clear that $\boldsymbol{\beta}_{\mathbf{x}}(\cdot)$ depends on θ. Then the function (the 'energy of SFD'):

$$C(\theta) = E\left\{\|\beta_{\mathbf{x}}(\mathbf{x})\|^2\right\} = \int_{\mathbf{x}} \|\beta_{\mathbf{x}}(\mathbf{x})\|^2 p_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x} \tag{12}$$

gives a complete description of the behavior of the SFD. By a "complete description of the behavior" we mean showing the behavior for different degrees of statistical dependence. For example, for $\theta = 0$, x_1 and x_2 are independent, and hence the SFD is zero and $C(0) = 0$. When θ increases, x_1 and x_2 will be dependent and hence the value of $C(\theta)$ will no longer be zero. Then when θ reaches $\frac{\pi}{2}$, x_1 and x_2 will become again independent, and SFD vanishes.

Consequently, we use the theoretically calculated $C(\theta)$ (for $\theta = 0$ to $\theta = \frac{\pi}{2}$) as a basis for comparison of the performances of different SFD estimators: the one which follows better the variation of $C(\theta)$ versus θ gives a better SFD estimation for different degrees of independence between random variables.

In the following two subsections, the theoretical values of the mutual information of x_1 and x_2 (as a measure of their dependence) and $C(\theta)$ versus θ will be calculated.

2.1 Mutual Information of random variables x_1 and x_2

To calculate the mutual information of x_1 and x_2, we first calculate their joint and marginal PDF's, and then use the definition given in (1).

Because of the independence of the random variables s_1 and s_2, the density of the random vector $\mathbf{s} \triangleq (s_1,s_2)^T$ is written as $p_{\mathbf{s}}(\mathbf{s}) = p_{s_1}(s_1)\, p_{s_2}(s_2)$, where $p_{s_i}(\cdot)$ is given by (9).

Now let $\mathbf{R} = \mathbf{R}(\theta)$ denote the mixing matrix in (11), which is the rotation matrix by angle θ, and $\mathbf{x} \triangleq (x_1,x_2)^T$. Then (11) is written as $\mathbf{x} = \mathbf{R}\mathbf{s}$, and hence:

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= p_s(\mathbf{R}^T \cdot \mathbf{x}) \\ &= p_{s_1}(x_1\cos\theta + x_2\sin\theta) \cdot p_{s_2}(-x_1\sin\theta + x_2\cos\theta) \end{aligned} \tag{13}$$

Consequently, the joint density of the random variables x_1 and x_2 is known. Now we calculate the marginal densities of these random variables.

Firstly, note that the sum of two random Gaussian variables with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 is a Gaussian random variable with mean $\mu_1+\mu_2$ and variance $\sigma_1^2+\sigma_2^2$. Consequently:

$$G(t;\mu_1,\sigma_1^2) * G(t;\mu_2,\sigma_2^2) = G(t;\mu_1+\mu_2,\sigma_1^2+\sigma_2^2) \tag{14}$$

where $G(t;\mu,\sigma^2)$ is defined in (10) and '*' stands for convolution. Moreover, if x is a Gaussian random variable with mean μ and variance σ^2, then ax (where a is a scalar) will be a Gaussian random variable with mean $a\mu$ and variance $a^2\sigma^2$. Consequently:

$$\frac{1}{|a|} G(\frac{t}{a};\mu,\sigma^2) = G(t;a\mu,a^2\sigma^2) \tag{15}$$

Now form $x_1 = s_1\cos\theta - s_2\sin\theta$, the PDF of x_1 is written as the convolution of the PDF's of random variables $s_1\cos\theta$ and $-s_2\sin\theta$. Moreover:

$$\begin{aligned} &p_{s_1\cos\theta}(t) = \frac{1}{|\cos\theta|} p_{s_1}(\frac{s_1}{\cos\theta}) = \\ &\frac{G(t;\cos\theta,\sigma^2\cos^2\theta) + G(t;-\cos\theta,\sigma^2\cos^2\theta)}{2} \end{aligned} \tag{16}$$

where the second equality is written by (9) and (14). In a similar manner:

$$p_{-s_2\sin\theta}(t) = \frac{G(t;-\cos\theta,\sigma^2\sin^2\theta) + G(t;\cos\theta,\sigma^2\sin^2\theta)}{2} \tag{17}$$

Convolving both sides of (16) and (17), and using (14) we obtain:

$$p_{x_1}(t) = \frac{1}{4}G(t;\mu_1,\sigma^2) + \frac{1}{4}G(t;-\mu_1,\sigma^2) + \frac{1}{4}G(t;\mu_2,\sigma^2) + \frac{1}{4}G(t;-\mu_2,\sigma^2) \quad (18)$$

where $\mu_1 \triangleq \cos\theta + \sin\theta$ and $\mu_2 \triangleq \cos\theta - \sin\theta$. Moreover, it can also be easily verified that x_2 has the same distribution given by (18).

Now it remains to do the integration of (1) to calculate the mutual information. We have done this integration numerically using MATLAB (for $\sigma = 0.5$), and obtained the diagram of Fig. 1.

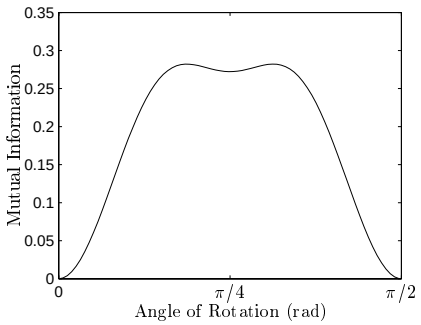

Fig. 1 Mutual Information of x_1 and x_2 versus θ

2.2 Score Function Difference of random vector x

For calculating the SFD, it suffices to calculate $\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x})$ and $\varphi_{\mathbf{x}}(\mathbf{x})$, and the use the definition in (3).

Combining (5) and (18), $\psi_i(\cdot)$, the i-th component of the MSF of $\mathbf{x}$, is obtained as:

$$\psi_i(t) = \frac{G'(t;\mu_1,\sigma) + G'(t;-\mu_1,\sigma) + G'(t;\mu_2,\sigma) + G'(t;-\mu_2,\sigma)}{G(t;\mu_1,\sigma) + G(t;-\mu_1,\sigma) + G(t;\mu_2,\sigma) + G(t;-\mu_2,\sigma)} \quad (19)$$

where $G'(t;\mu,\sigma)$ is the derivative of $G(t;\mu,\sigma)$, defined in (10), with respect to t.

From (1), for calculating the JSF od $\mathbf{x}$ we need the partial derivatives of $p_{\mathbf{x}}(\cdot)$ with respect to x_1 and x_2. This can be easily derived from (13):

$$\frac{\partial p_{\mathbf{x}}}{\partial x_1}(\mathbf{x}) = p'_{s_1}(s_1)p_{s_2}(s_2)\cos\theta - p_{s_1}(s_1)p'_{s_2}(s_2)\sin\theta \quad (20)$$

$$\frac{\partial p_{\mathbf{x}}}{\partial x_2}(\mathbf{x}) = p'_{s_1}(s_1)p_{s_2}(s_2)\sin\theta + p_{s_1}(s_1)p'_{s_2}(s_2)\cos\theta \quad (21)$$

where $s_1 = x_1\cos\theta + x_2\sin\theta$, and $s_2 = -x_1\sin\theta + x_2\cos\theta$. The above equations, completely determine $\boldsymbol{\beta}_{\mathbf{x}}(x_1,x_2)$ for each value of (x_1,x_2).

Finally, knowing the explicit expressions of all the functions of the integral in (12), it may be numerically calculated. Fig. 2 shows the obtained $C(\theta)$ versus θ, for $\sigma = 0.5$.

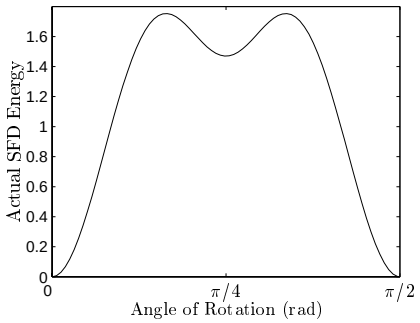

Fig. 2 $C(\theta)$, the energy of SFD, versus θ

2.3 Proposed Comparison Criterion

In the previous section, the exact value of $C(\theta)$ versus θ has been calculated. Now, for evaluating the performance of different SFD estimation methods, we create two source signals according to the distribution given by (9). Then, for each value of θ (in the range 0 to $\pi/2$), the observation samples are calculated using (11). The SFD of these observations are then estimated by different SFD estimation methods. Finally, for each SFD estimation algorithm, we plot the obtained $\hat{C}(\theta)$, and visually compare it with the true $C(\theta)$ of Fig. 2.

3 Experimental results

We have applied the explained method to the existing SFD estimators proposed in [6], [7], [8]. In all simulations, 500 data samples have been used for estimating the SFD. The simulation has been done 100 times for each method, and the obtained $\hat{C}(\theta)$ have been averaged through these 100 simulations. Fig.'s 3 through 7 show the resulted averaged $C(\theta)$ for each estimator versus θ.

Remark 1. Fig.'s 3 through 7 clearly show the performance of different SFD estimation, in the sense of their ability to *follow* the exact variations of the energy of SFD, as it is seen in Fig. 2. However, for none of these estimators the 'value' of the estimated SFD is correct. If we re-run the simulations with a different number of sample points (instead of 500, as in the above figures), we will obtain different 'values' for the

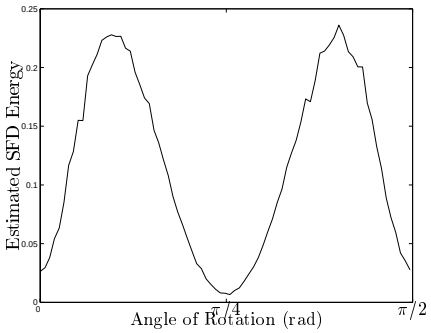

Fig. 3 Polynomial method.

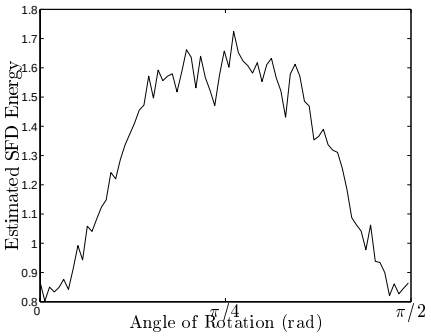

Fig. 4 Histogram method.

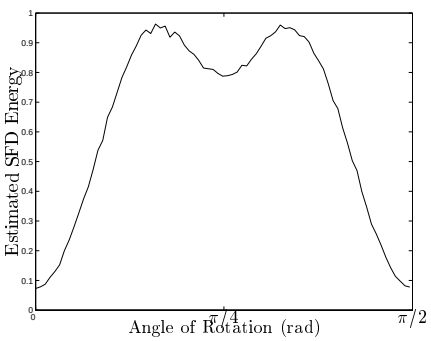

Fig. 5 Kernel method.

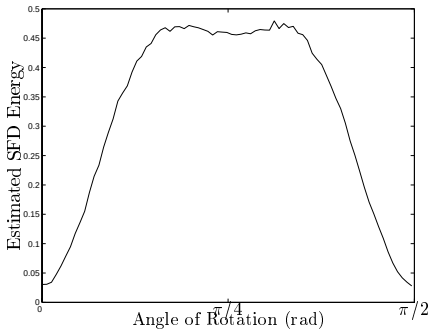

Fig. 6 Pham's method.

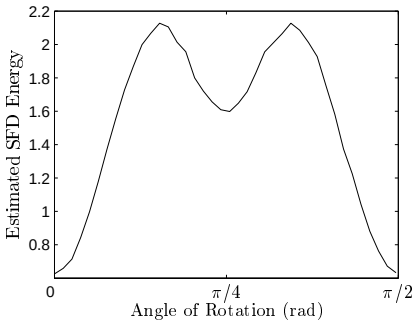

Fig. **7** The authors' method ([8]).

estimated SFD's. This shows that there is a kind of 'normalization' in the current estimators of SFD (that is, the estimated values depend on the number of data points). This problem has not been already noticed in BSS algorithms, because they used SFD (as the gradient of mutual information) in gradient based algorithms (e.g. $\mathbf{y} \leftarrow \mathbf{y} - \mu\beta_{\mathbf{y}}$ for the MP approach [6]), and any scaling error in the estimation of SFD is absorbed within the step size parameter of these algorithms (μ).

Remark 2. As stated before, the quality of a SFD estimator (in the sense of this paper) does not generally insure a good quality of a BSS algorithm based on that estimator. In fact, for having a good separation quality in a BSS algorithm, we just need that the SFD be well estimated for *nearly independent* random variables (around $\theta = 0$ and $\theta = \pi/2$ in the above figures). This is because a gradient based BSS algorithm converges when the estimated SFD vanishes. If the estimated SFD vanishes in the correct place (independent outputs), we obtain a good separation quality. For dependent signals ($0 < \theta < \pi/2$ in the above figures), the errors in the values of the estimated SFD are not so important for a BSS algorithm, provided that these errors do not 'badly' change the sign of the estimated derivative of MI of outputs with respect to the parameters of the separating system.

A good example is estimating the inverse of the system (11) by the algorithm $\theta \leftarrow \theta - \mu\partial I(\mathbf{Rx})/\partial\theta$, and using the polynomial SFD estimator. Then it is shown [9] that the poor behavior of this estimator around $\theta = \pi/4$ (see Fig. 3), will even change the sign of the estimated $\partial I(\mathbf{Rx})/\partial\theta$, but in an advantageous manner: it prevents the algorithm from getting trapped in the local minimum of MI at $\theta = \pi/4$ (see Fig. 1).

Conclusions

In this paper, the SFD for a special kind of distribution was first theoretically calculated. Then, this actual value of SFD was used as a basis for comparing the performance of the existing SFD estimators. We found out that a normalization error exists in all of the existing SFD estimators, which

has not been already seen because of the manner these estimators are used in BSS algorithms. The reason of this error has not yet been known, and needs more investigation.

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", *Signal Processing*, vol. 85, no. 5, pp. 975–995, May 2005.

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

[**8**] B. Bahmani, M. Babaie-Zadeh, and C. Jutten, "A new method for estimating Score Function Difference (SFD) and it's application to blind source separation", To appear in *EUSIPCO 2005*.

[**9**] M. Babaie-Zadeh, B. Bahmani, and C. Jutten, "ICA by mutual information minimization: An approach for avoiding local minima", To appear in *EUSIPCO 2005*.

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

[**11**] 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.