ICA BY MUTUAL INFORMATION MINIMIZATION: AN APPROACH FOR AVOIDING LOCAL MINIMA

Massoud Babaie-Zadeh[1]*, Bahman Bahmani*[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.
mbzadeh@yahoo.com, bahmanibahman@yahoo.com, Christian.Jutten@inpg.fr

ABSTRACT

Using Mutual Information (MI) minimization is very common in Blind Source Separation (BSS). However, it is known that gradient descent approaches may trap in local minima of MI in constrained models. In this paper, it is proposed that this problem may be solved using a 'poor' estimation of the derivative of MI.

1. INTRODUCTION

Blind Source Separation (BSS) is the problem of retrieving some statistically independent source signals from mixtures of them, when there is no information about the sources or about the mixture. This problem had been extensively under study since mid 80's[1], and there are currently quite a lot of algorithms for solving it (see for example [2] and [3]).

Let $\mathbf{s} = (s_1, \ldots, s_N)^T$ and $\mathbf{x} = (x_1, \ldots, x_N)^T$ be the vector of sources and observations, respectively. Then for linear instantaneous mixtures, $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{A}$ is the unknown mixing matrix. Since the sole information about the source signals is their statistical independence, one approach for source separation is estimating a separating matrix $\mathbf{B}$, which transforms again, through $\mathbf{y} = \mathbf{B}\mathbf{x}$, the observations into statistically independent outputs ($\mathbf{y}$). Hence the method name: Independent Component Analysis (ICA). It is well known [4] that this transformation results in source separation up to trivial indeterminacies of scale and permutation of sources.

In BSS, the independence of the outputs cannot be simplified to decorrelation (second-order independence). In other words, output decorrelation (which is usually called Principal Component Analysis or PCA) does not insure source separation. However, it is well known [5] that PCA leaves just a 'rotation' to be estimated. Dividing the ICA task into these two stages (*i.e.* PCA or prewhitening and then a rotation) has been used in many ICA algorithms [2, 5].

Consider now the two sources and two sensors case, and suppose that the prewhitening has been already done. Then the outputs y_1 and y_2 are related to the sources by:

$$\begin{cases} y_1 = \cos(\theta)s_1 - \sin(\theta)s_2 \\ y_2 = \sin(\theta)s_1 + \cos(\theta)s_2 \end{cases} \tag{1}$$

For solving the ICA problem, the angle θ which results in independent outputs has to be estimated.

One approach to measure the statistical independence of outputs is to use their mutual information $I(\mathbf{y})$. Then, the separation algorithm is based on using gradient based approaches for minimizing $I(\mathbf{y})$. If, for simplicity of notations, and also for explicitly indicating the dependence of $I(\mathbf{y})$ to θ, we define $F(\theta) \triangleq I(\mathbf{y})$ and $f(\theta) \triangleq F'(\theta) = dI(\mathbf{y})/d\theta$, then the source separation algorithm is:

$$\theta \leftarrow \theta - \mu \frac{dI(\mathbf{y})}{d\theta} \quad \text{or} \quad \theta \leftarrow \theta - \mu f(\theta) \tag{2}$$

Minimizing mutual information (MI) of outputs (using gradient based methods) has been used in many ICA algorithms for different kinds of mixtures [6, 7, 8, 4]. This approach has several advantages over other BSS approaches, including: 1) It is shown [5] that for linear instantaneous mixtures it results in an asymptotically Maximum Likelihood (ML) estimation of source signals; 2) Contrary to some independence criteria (like 4th order cross-cumulants) it has no approximation, that is, it vanishes if and only if the outputs are statistically independent. Consequently, it can be used for separating more complicated mixtures (*e.g.* non-linear mixtures); 3) It may result in a *unifying* approach for separating different kinds of separable models [6].

Although it is shown that Mutual Information has no local 'minima' [6], it is seen that in a constrained model like (1), $I(\mathbf{y})$ has local minima with respect to θ [9]. Consequently, without some precautions, the ICA algorithms based on mutual information minimization by steepest descent approaches are not reliable.

In this paper, we are going to present an approach for avoiding local minima in (2) by using 'poor' estimation of $dI(\mathbf{y})/d\theta$. The main idea is that in practice, we cannot use the algorithm (2), because $f(\theta) = dI(\mathbf{y})/d\theta$ is not known and it must be estimated from the data. Consequently, the practical algorithm is:

$$\theta \leftarrow \theta - \mu \hat{f}(\theta) \tag{3}$$

where $\hat{f}(\theta)$ is an estimation of $dI/d\theta$. Then, *although we know that (2) has local minima, it is possible that (3) has no local minima, depending on the estimation method of* $\frac{dI}{d\theta}$.

To summarize, we are going to present in this paper, an estimation method for $f(\theta) = dI(\mathbf{y})d\theta$, which does not result in a accurate estimation of $f(\theta)$, but results in an algorithm (3) for which is guaranteed to have no local minima.

2. PRELIMINARY ISSUES

2.1 Gradient of Mutual Information

The following theorem can be easily proved by using a general expression for the gradient of MI [10]. However, be-

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).

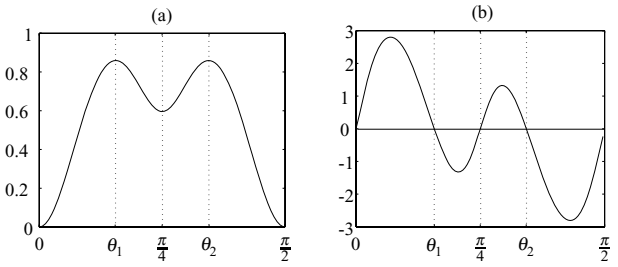

Figure 1: a) Mutual Information versus θ, b) Its derivative with respect to θ.

cause of lack of space, the proof has been omitted here.

Theorem 1 *In model (1), the gradient of $I(\mathbf{y})$ with respect to θ is given by:*

$$f(\theta) \triangleq \frac{dI(\mathbf{y})}{d\theta} = E\left\{y_1\psi_2(y_2) - y_2\psi_1(y_1)\right\} \tag{4}$$

where $\psi_i(y_i)$ is the score function of y_i, defined by:

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

in which $p_{y_i}(y_i)$ stands for the Probability Distribution Function (PDF) of y_i.

2.2 Local minima of the algorithm (2)

We know that $F(\theta)$, and hence the algorithm (2) may contain local minima [9]. Consider first the following example:

Example 1. Let the sources s_1 and s_2 have bi-modal Gaussian density:

$$p_{s_1}(s) = p_{s_2}(s) = \frac{1}{2}\left\{N(s;1,0.3) + N(s;-1,0.3)\right\} \tag{6}$$

where $N(x;\mu,\sigma) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{\frac{(x-\mu)^2}{2\sigma^2}\}$ represents the Gaussian density with mean μ and variance σ^2. Then, from (1) the PDF's of y_1 and y_2 can be obtained, and from there $F(\theta) = I(\mathbf{y})$ and its derivative ($f(\theta)$) can be theoretically computed. Figure 1 shows the plot of F and f versus θ for $0 \le \theta \le \frac{\pi}{2}$. It is seen in this figure that $F(\theta)$ has a local minimum at $\theta = \frac{\pi}{4}$.

Figure 1.b shows that $f(\theta)$ has 3 zeros in the interval $0 < \theta < \frac{\pi}{2}$. However, it is important to note that all these zeros are not the *stationary* (or *stable*) points of the algorithm (2), because $f'(\theta)$ is negative at $\theta = \theta_1$ and $\theta = \theta_2$. Consequently, a small deviation in θ around the point $\theta = \theta_1$ (or $\theta = \theta_2$) in algorithm (2) results in moving in the same direction, and going farther. In fact, $\theta = \theta_1$ and $\theta = \theta_2$ correspond to local *maxima* of $F(\theta)$, and not local minima. On the other hand, $\theta = \frac{\pi}{4}$ is a stationary point of the algorithm (2) because $f(\frac{\pi}{4}) = 0$ and $f'(\frac{\pi}{4}) > 0$.

The above example clarifies the fact that *the stationary points of the algorithm (2) are the points for which $f(\theta) = 0$ and $f'(\theta) > 0$.*

2.3 Estimating the gradient of MI

Unlike example 1, the distributions of the sources are not usually known in BSS. Consequently, $f(\theta)$ cannot be computed exactly using (4) and it must be estimated from the data (*i.e.* output samples). From Theorem 1, $f(\theta)$ can be estimated by:

$$\hat{f}(\theta) = \hat{E}\left\{y_1\hat{\psi}_2(y_2) - y_2\hat{\psi}_1(y_1)\right\} \tag{7}$$

where $\hat{\psi}_i(y_i)$ is an estimation of the score function of y_i, and $\hat{E}$ stands for the estimation of the expected value by averaging throughout all data samples.

There are several methods for estimating score functions already used in BSS literature. Consider for example the following estimation methods:

Histogram estimation. In this approach, the output PDF's are first estimated by a simple histogram. Then, approximating the derivative in (5) by a difference, a simple histogram estimation for ψ_i is obtained.

Kernel estimation. Having observations $\{y_1, y_2, \ldots, y_N\}$ from a random variable y, the kernel estimation of its PDF is given by [11]:

$$\hat{p}_y(y) = \frac{1}{N}\sum_{k=1}^{N} K\left(\frac{y - y_k}{h}\right) \tag{8}$$

where $K(\cdot)$ is a 'kernel' (*i.e.* the PDF of a random variable with zero mean and unit variance), and h is the bandwidth of the estimator (a too small bandwidth results in a very 'noisy' estimated PDF, a too large bandwidth results in a PDF which is roughly the same as the kernel itself). Using this estimator for PDF of y, the score function is estimated by $\hat{\psi}_y(y) = -\hat{p}'_y(y)/\hat{p}_y(y)$.

Polynomial estimation. It is well known [12] that under very mild conditions, for a function $f(\cdot)$, we have:

$$E\left\{f(y)\psi_y(y)\right\} = E\left\{f'(y)\right\} \tag{9}$$

where $\psi_y(\cdot)$ is the score function of the random variable y. This equation provides a basis to design Minimum Mean Square Error (MMSE) estimators for ψ_y. Let for example $\psi_y(y)$ be estimated as a linear combination of the functions $k_1(y), k_2(y), \ldots, k_N(y)$:

$$\hat{\psi}_y(y) = w_1k_1(y) + w_2k_2(y) + \cdots + w_Nk_N(y) = \mathbf{k}^T(y)\mathbf{w} \tag{10}$$

where $\mathbf{w} \triangleq (w_1, \ldots, w_N)^T$ and $\mathbf{k}(y) \triangleq (k_1(y), \ldots, k_N(y))$. In this equation, $\mathbf{w}$ must be determined such that the mean square error $E\left\{(\psi_y(y) - \hat{\psi}_y(y))^2\right\}$ be minimized. From the principal of orthogonality [13], $E\left\{\mathbf{k}(y)\left(\psi_y(y) - \hat{\psi}_y(y)\right)\right\}$, which using (9) becomes:

$$E\left\{\mathbf{k}(y)\mathbf{k}^T(y)\right\}\mathbf{w} = E\left\{\mathbf{k}'(y)\right\} \tag{11}$$

This equation determines the optimum $\mathbf{w}$, without the need of knowing ψ_y. For example, if we choose $k_1(y) = 1$, $k_2(y) = y$, $k_3(y) = y^2$ and $k_4(y) = y^3$, we will have a 3rd order polynomial estimation of the score function. For a symmetric random variable y, $p_y(y)$ is an even function, and hence $\psi_y(y)$ is odd. Consequently, we can choose $k_1(y) = y$ and $k_2(y) = y^3$, which results in the following estimator for ψ_y:

$$\hat{\psi}_y(y) = w_1y + w_2y^3. \tag{12}$$

Solving (11) for this simple case, the coefficients w_1 and w_2 of the above estimator are given by:

$$w_1 = \frac{Ey^6 - 3(Ey^4)(Ey^2)}{(Ey^2)(Ey^6) - (Ey^4)^2}, \quad w_2 = \frac{3(Ey^2)^2 - Ey^4}{(Ey^2)(Ey^6) - (Ey^4)^2} \tag{13}$$

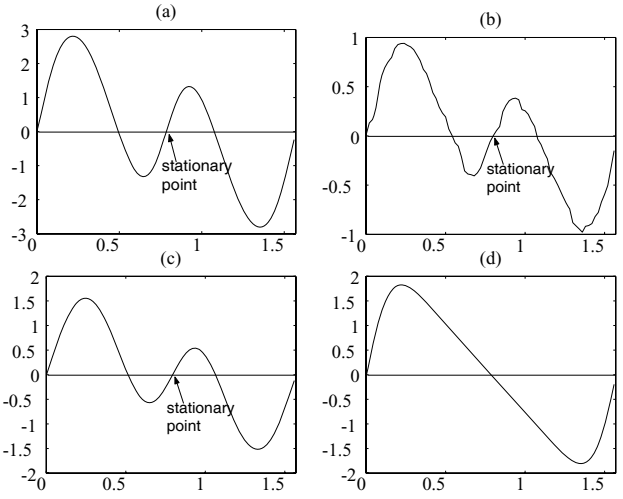

Figure 2: a) Theoretical derivative of MI with respect to θ, b) Its histogram estimation, c) Its kernel estimation, d) Its polynomial estimation.

3. THE MAIN IDEA

The main idea of the paper comes from the following experiment:

Experiment. For the sources of Example 1, and for y_1 and y_2 defined by (1), we have estimated $f(\theta) = dI(\mathbf{y})/d\theta$ from (7) and the score functions (using 100 data samples) using the three estimators of Section 2.3. The result is shown in Fig. 2. As it is seen in this figure, kernel estimator provides a better estimation of score function compared to the others.

Now consider the polynomial estimator. At a first glance, one may decide to not use this estimator: it has not enough degree of freedom (since it is based on a 3rd order polynomial) to follow the variations of $f(\theta)$, and hence it does not provide a very satisfactory estimation of it. However, a closer look at this estimator, reveals a surprising property: because of the limited degree of freedom, $\hat{f}(\theta)$ has just one zero in the interval $0 < \theta < \frac{\pi}{2}$, and *the slope of $\hat{f}(\theta)$ at this point is negative, consequently it is not a stationary point of the algorithm (3)*.

To summarize, although polynomial estimator does not provide a very good estimation of $dI/d\theta$, it results in a local minimum free version of the algorithm (3). Recall that our main goal, too, was not obtaining a good estimation of $dI/d\theta$, it was separating the sources. In fact, here, this 'poor' estimation of $dI/d\theta$ is highly better, since it automatically solves the problem of local minima.

This leads us to the following idea: *use the algorithm (3), along with polynomial estimation of $f(\theta)$. Then the separation algorithm is free of local minima*.

However, Fig. 2 is only for the source distribution of Example 1. Does the above result remain valid for any source distribution? The next section states that for the case in which the sources have the same distribution, the answer is positive.

4. THE MAIN THEOREM

In Section 2.3, we saw that for estimating the score function of a symmetric random variable using 3rd order polynomials, we can use polynomials of the form $w_1 y + w_2 y^3$. This is justified by the fact that for a symmetric random variable, $p_y(y)$ is even, and hence $\psi_y(y)$ is odd.

However, here we propose to use the same estimator, *i.e.* the polynomials of the form $w_1 y + w_2 y^3$, for estimating the score functions of the outputs of the system (1), either for symmetric sources or for asymmetric sources. Note that for asymmetric sources, $\psi_i(.)$ is no more odd, and hence estimating it by a polynomial of the form $w_1 y_i + w_2 y_i^3$ does not provide a 'good' estimation of the score function. But, we are not seeking a 'good' estimation of score functions, our main objective is to have a source separation algorithm. The next theorem, which is the main theorem of the paper, shows that such an approach provide a source separation algorithm (based on MI minimization) which is free of local minima.

Theorem 2 *If the sources s_1 and s_2 have the same distribution (and non-binary), and if the output score functions are estimated using polynomials of the form $w_1 y + w_2 y^3$, and are applied for estimating $f(\theta)$ from (7), then the algorithm (3) is free of local minima.*

To prove this theorem, we first need to define the following notations.

Definition 1 *For a zero mean random variable y, we define:*

a) $\kappa_y = E\{y^4\} - 3E\{y^2\}^2$.

b) $\lambda_y = E\{y^2\}E\{y^6\} - E\{y^4\}^2$.

Note that κ_y is in fact the 4th-order cumulant of y.

Lemma 1 *For any random variable y, $\lambda_y \geq 0$. Moreover, the equality holds only for binary random variables.*

Proof: The Cauchy inequality implies that for any random variables z and t:

$$E\{z^2\}E\{t^2\} \geq E\{zt\}^2 \tag{14}$$

where the equality holds only where $t = kz$ (for a constant k). For $z = y$ and $t = y^3$, (14) becomes $\lambda_y \geq 0$. Moreover, the equality holds only for the case $y^3 = ky$, which implies $y^2 = k$, that is, where y is a binary random variable. □

Proof of Theorem 2: Let $p_k \triangleq E\{y_1^k\}$ and $q_k \triangleq E\{y_2^k\}$ denote the k-th order moments of y_1 and y_2, respectively. Then from (1) and using the fact $E\left\{s_1^{k_1} s_2^{k_2}\right\} = 0$, where $k_1 = 1$ or $k_2 = 1$, and after doing a few calculations we obtain:

$$\begin{cases} p_2 = q_2 = m_2 \\ p_4 = q_4 = \frac{1}{2}\left(3m_2^2 - m_4\right)\sin^2(2\theta) + m_4 \\ p_6 = -\frac{5}{2}m_3^2 \sin^3(2\theta) + \frac{3}{4}\left(5m_4 m_2 - m_6\right)\sin^2(2\theta) + m_6 \\ q_6 = \frac{5}{2}m_3^2 \sin^3(2\theta) + \frac{3}{4}\left(5m_4 m_2 - m_6\right)\sin^2(2\theta) + m_6 \end{cases} \tag{15}$$

Now, let ψ_1 and ψ_2, the score functions of y_1 and y_2, be estimated as $\hat{\psi}_1(y_1) = v_1 y_1 + v_2 y_1^3$ and $\hat{\psi}_2(y_2) = w_1 y_1 + w_2 y_1^3$. Combining these equations with (1) and (4), and after doing some calculation, we obtain:

$$\hat{f}(\theta) = -\frac{1}{4}\kappa(v_2 + w_2)\sin(4\theta) \tag{16}$$

where $\kappa \triangleq \kappa_{s_1} = \kappa_{s_2} = m_4 - 3m_2^2$. From (13), the optimum values for v_2 and w_2 are:

$$v_2 = \frac{3p_2^2 - p_4}{p_2 p_6 - p_4^2} = -\frac{\kappa_{y_1}}{\lambda_{y_1}}, \quad w_2 = \frac{3q_2^2 - q_4}{q_2 q_6 - q_4^2} = -\frac{\kappa_{y_2}}{\lambda_{y_2}} \tag{17}$$

Using these values, (16) is written as:

$$\hat{f}(\theta) = \frac{1}{4}\kappa\left(\frac{\kappa_{y_1}}{\lambda_{y_1}} + \frac{\kappa_{y_2}}{\lambda_{y_2}}\right)\sin(4\theta) \tag{18}$$

Now, from (15) and after some calculations, we obtain $\kappa_{y_1} = \kappa_{y_2} \triangleq p_4 - 3p_2^2 = \frac{1}{2}\left(2 - \sin^2(2\theta)\right)\kappa$, consequently:

$$\hat{f}(\theta) = \frac{\kappa^2\left(2 - \sin^2(2\theta)\right)\left(\lambda_{y_1} + \lambda_{y_2}\right)\sin(4\theta)}{8\lambda_{y_1}\lambda_{y_2}} \tag{19}$$

Also, from (15):

$$\lambda_{y_1} \triangleq -\frac{\kappa^2}{4}\sin^4(2\theta) - \frac{5}{2}m_2m_3^2\sin^3(2\theta) + \left(\frac{3}{4}m_4m_2^2 - \frac{3}{4}m_2m_6 + m_4^2\right)\sin^2(2\theta) + m_2m_6 - m_4^2 \tag{20a}$$

$$\lambda_{y_2} = -\frac{\kappa^2}{4}\sin^4(2\theta) + \frac{5}{2}m_2m_3^2\sin^3(2\theta) + \left(\frac{3}{4}m_4m_2^2 - \frac{3}{4}m_2m_6 + m_4^2\right)\sin^2(2\theta) + m_2m_6 - m_4^2 \tag{20b}$$

Since the sources are assumed to be non-binary, $\lambda_{y_1} + \lambda_{y_2} > 0$. Moreover, $2 - \sin^2(2\theta)$ is always positive. Consequently, from (19), $\hat{f}(\theta) = 0$ if and only if $\sin(4\theta) = 0$. Therefore, the only zeros of $\hat{f}(\theta)$ in the interval $0 \le \theta \le \frac{\pi}{2}$ are $\theta = 0$, $\theta = \pi/2$, and $\theta = \pi/4$. Substituting (20) in (19) and doing some tedious calculations, it is obtained:

$$\hat{f}'(0) = \hat{f}'(\frac{\pi}{2}) = 2\frac{(m_4 - 3m_2^2)^2}{m_2m_6 - m_4^2} = 2\frac{\kappa^2}{\lambda} \tag{21}$$

where $\lambda \triangleq m_2m_6 - m_4^2$, and:

$$\hat{f}'(\frac{\pi}{4}) = \frac{\kappa^2(m_4^2 - 9m_2^2m_4 + 9m_2^4 - m_2m_6)}{4\left(\lambda_{y_1}\big|_{\theta=\frac{\pi}{4}}\right)\left(\lambda_{y_2}\big|_{\theta=\frac{\pi}{4}}\right)} \tag{22}$$

From (21), $\hat{f}'(0) > 0$ and $\hat{f}'(\frac{\pi}{2}) > 0$. Consequently, $\theta = 0$ and $\theta = \pi/2$ are the only stationary points of the algorithm (3), which result in source separation.

To show that $\theta = \pi/4$ is not an stationary point of this algorithm, from Section 2.2, we must show that $\hat{f}'(\frac{\pi}{4}) < 0$. From (22), the sign of $\hat{f}'(\frac{\pi}{4})$ is the same as $m_4^2 - 9m_2^2m_4 + 9m_2^4 - m_2m_6$, which can be written as $-(\lambda + 9m_2^2(m_4 - m_2^2))$. However, writing the Cauchy inequality (14) for $z = s^2$ and $t = 1$, shows that $m_4 \ge m_2^2$. Consequently, $-(\lambda + 9m_2^2(m_4 - m_2^2)) < 0$, which proves the theorem. □

Remark 1. In Theorem 2 and its proof, it is implicitly assumed that there are enough data samples to insure a very good estimation of the expectation operation ($E\{\}$).

Remark 2. Note that in the proposed approach (applying 3rd order estimation of score functions for BSS), we have excluded binary sources. In fact, the score function of a binary random variable cannot be estimated as (12), because from (13) and Lemma 1, the coefficients of such an estimator will be infinity.

Remark 3. From (19), if $\kappa = 0$ then $\hat{f}'(\theta) = 0$, $\forall\theta$. This is compatible with the well known fact that BSS is not possible for Gaussian sources (for which 4-th order cumulants vanish).

5. CONCLUSION

In this paper, we showed that a 'poor' estimation of the gradient may have advantages in a steepest descent gradient algorithm. Indeed, when BSS is achieved by a MI minimization algorithm, a poor estimation of score functions based on 3rd order polynomials avoids the existence of local minima.

The main theorem of the paper (Theorem 2) was stated and proved just for two sources with identical distributions. When the distribution of the sources are different, we have by now, neither a proof, nor a counter-example. This will restrict the usefulness of the result of the paper. However, the main point of the paper is the possibility of solving the problem of local minima through 'poor' estimation of the gradient of the cost function. The extension of the result to more general cases is currently under study.

REFERENCES

[1] J. Hérault and C. Jutten, "Space or time adaptive signal processing by neural networks models", in *Intern. Conf. on Neural Networks for Computing*, Snowbird (Utah, USA), 1986, pp. 206–211.

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

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

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

[5] J.-F. Cardoso, "Blind signal separation: statistical principles", *Proceedings IEEE*, vol. 9, pp. 2009–2025, 1998.

[6] M. Babaie-Zadeh and C. Jutten, "A general approach for mutual information minimization and its application to blind source separation", vol. 85, no. 5, pp. 975–995, May 2005.

[7] 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.

[8] A. Taleb and C. Jutten, "Entropy optimization, application to blind source separation", in *Proceedings of ICANN'97*, Lausanne, Switzerland, October 1997, pp. 529–534.

[9] F. Vrins and M. Verleysen, "On the entropy minimization of a linear mixture of variables for source separation", vol. 85, no. 5, pp. 1031–1046, May 2005.

[10] 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.

[11] B. W. Silverman, *Density estimation for statistics and data analysis*, Chapman and Hamm, 1986.

[12] D. D. Cox, "A penalty method for nonparametric estimation of the logarithmic derivative of a density function", *Ann. Instit. Statist. Math.*, vol. 37, pp. 271–288, 1985.

[13] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 2002.