

Independent Component Analysis Based on Nonparametric Density Estimation

Riccardo Boscolo, Hong Pan, *Member, IEEE*, and Vwani P. Roychowdhury

Abstract—In this paper, we introduce a novel independent component analysis (ICA) algorithm, which is truly blind to the particular underlying distribution of the mixed signals. Using a nonparametric kernel density estimation technique, the algorithm performs simultaneously the estimation of the unknown probability density functions of the source signals and the estimation of the unmixing matrix. Following the proposed approach, the blind signal separation framework can be posed as a nonlinear optimization problem, where a closed form expression of the cost function is available, and only the elements of the unmixing matrix appear as unknowns. We conducted a series of Monte Carlo simulations, involving linear mixtures of various source signals with different statistical characteristics and sample sizes. The new algorithm not only consistently outperformed all state-of-the-art ICA methods, but also demonstrated the following properties: 1) Only a flexible model, capable of learning the source statistics, can consistently achieve an accurate separation of all the mixed signals. 2) Adopting a suitably designed optimization framework, it is possible to derive a flexible ICA algorithm that matches the stability and convergence properties of conventional algorithms. 3) A nonparametric approach does not necessarily require large sample sizes in order to outperform methods with fixed or partially adaptive contrast functions.

Index Terms—Independent component analysis (ICA), kernel density estimation, nonlinear optimization, nonparametric methods.

I. INTRODUCTION

IN recent years, Independent Component Analysis (ICA) algorithms have proven successful in separating linear mixtures of independent source signals [1]–[12]. While most of the existing implementations have been tested and compared to each other using synthetic data, significant results on separating real world mixtures of signals have been reported as well [13]–[18]. Many existing methods rely on simple assumptions on the source statistics and are characterized by well assessed convergence and consistency properties [19]. When such hypotheses hold strictly or are only moderately violated, most conventional ICA algorithms are capable of quickly and efficiently achieve the desired source separation. However, such algorithms can perform suboptimally or even fail to produce the desired

source separation, when the assumed statistical model is inaccurate [5].

A relevant example and a well-known ICA implementation is Hyvärinen's FastIca [9], which requires the user to select a contrast function according to the hypothetical (but unknown) probability density functions (pdfs) of the sources to be reconstructed. Such issues do not arise in the case of moment based implementations of blind signal separation algorithm (e.g., Cardoso's Jade [20]). However, these approaches usually rely exclusively on third or fourth order cross-cumulants in order to measure independency, and represent just an approximation of the mutual information minimization principle [6]. Clearly, when the separation of signals from real world data is attempted, such constraints are highly undesirable.

Alternative methods that employ a more flexible model for the pdf of the source signals have been introduced [21]–[23]. These methods usually consist of a parametric density estimation technique that alternates with a cost function optimization step in an iterative approximation framework. Although these approaches tend to outperform standard algorithms in specific cases (e.g., skewed sources), neither their convergence properties, nor their capability of modeling arbitrarily distributed sources, have been fully assessed. The recent introduction of kernel-based methods, such as Bach and Jordan's [24], demonstrate that finding a compromise between computational complexity, performance and strong convergence properties in a blind signal separation framework is still an open and challenging problem.

In this paper, we recognize the importance of defining a signal separation algorithm that is truly "blind" to the particular underlying distributions of the mixed signals, especially when real world applications are sought. A novel nonparametric ICA algorithm is introduced, which simultaneously estimates the unknown probability density functions of the source signals and the linear operator that allows the separation of the mixed signals (the so-called "unmixing matrix"). The resulting algorithm is nonparametric, data-driven, and does not require the definition of a specific model for the density functions.

The theoretical framework for the method we are proposing is derived in Section II, after a brief review of the conventional ICA separation principle, based on the minimization of the mutual information between the reconstructed signals. The key issues related to the actual algorithmic implementation of the proposed technique are addressed in Section III. In particular, the problem of local versus global convergence is investigated, and conditions ensuring the convergence of the proposed algorithm to the global optimum are suggested, for the case of mixtures of two signals. An extensive set of simulation experiments were

Manuscript received April 16, 2002; revised March 18, 2003. The work of R. Boscolo and V. P. Roychowdhury was supported in part by a Grant from the BioSpice program of DARPA, under Contract F30602-01-2-0557.

R. Boscolo and V. P. Roychowdhury are with the Electrical Engineering Department, University of California, Los Angeles, CA 90095 USA (e-mail: riccardo@ee.ucla.edu; vwani@ee.ucla.edu).

H. Pan is with the Functional Neuroimaging Laboratory, Department of Psychiatry, Weill Medical College, Cornell University, New York, NY 10021 USA (e-mail: hop2001@med.cornell.edu).

Digital Object Identifier 10.1109/TNN.2003.820667

conducted in order to demonstrate the performance improvement obtained with the proposed technique, when compared to other state-of-the-art ICA algorithms (Section IV).

II. JOINT ESTIMATION OF THE UNMIXING MATRIX AND OF THE DISTRIBUTION OF THE SOURCE SIGNALS

A. ICA Model and Separation Principle

The conventional generative model is assumed, where N independent and stationary source signals s_1, \dots, s_N are mixed by an unknown, full-rank mixing matrix A (size $N \times N$), resulting in a set of mixtures given by $\mathbf{x} = A\mathbf{s}$. The reconstruction of the original sources is attempted through a linear projection of the type $\mathbf{y} = W\mathbf{x}$ (W is also $N \times N$), with the assumption that at the most one of the sources has a Gaussian density [1]. The basic principle behind most ICA frameworks is the minimization of the mutual information between the reconstructed signals [25], that is

$$W_{\text{opt}} = \arg \min_W I(y_1, \dots, y_N). \quad (1)$$

This principle is characterized by having the minimum asymptotic variance, as shown by Donoho in [26], and it is also equivalent to the maximum likelihood (ML) principle when the source distributions are known [5], [27]. Using basic information theory equalities [28], expression (1) can be written as

$$\min_W \left\{ \sum_{i=1}^N H(y_i) - \log |\det W| - H(\mathbf{x}) \right\}. \quad (2)$$

Since the term $H(\mathbf{x})$ is a constant with respect to W , the objective function is reduced to

$$L(W) = \sum_{i=1}^N H(y_i) - \log |\det W| \quad (3)$$

$$= - \sum_{i=1}^N E [\log p_{y_i}(\mathbf{w}_i \mathbf{x})] - \log |\det W| \quad (4)$$

where \mathbf{w}_i is the i th row of the matrix W .

B. Nonparametric Kernel Density Estimation

In order to evaluate the marginal entropies $H(y_i)$ in (3), a model for the distribution of the unknown signals is required. In a quite effective way, Cardoso shows in [5], that incorrect assumptions on such distributions can result in poor estimation performance, sometimes in a complete failure to obtain the source separation.

To tackle this issue, we propose a nonparametric model, where the probability density functions p_{y_i} are directly estimated from the data using a kernel density estimation technique [29], [30]. The proposed approach allows a direct evaluation of the cost function and its derivatives, thus lifting the requirement of separating the optimization step from the step involving the re-estimation of the score functions, as in [21] or [23]. Given a

batch of sample data of size M , the marginal distribution of an arbitrary reconstructed signal is approximated as follows:

$$p_{y_i}(y_i) = \frac{1}{Mh} \sum_{m=1}^M \phi \left(\frac{y_i - Y_{im}}{h} \right), \quad i = 1, \dots, N \quad (5)$$

where h is the kernel bandwidth and ϕ is the Gaussian kernel

$$\phi(u) \triangleq \frac{1}{\sqrt{2\pi}} e^{-u^2/2}. \quad (6)$$

The kernel centroids Y_{im} are equal to

$$Y_{im} = \mathbf{w}_i \mathbf{x}^{(m)} = \sum_{n=1}^N w_{in} X_{nm} \quad (7)$$

where $\mathbf{x}^{(m)}$ is the m th column of the mixture matrix X . This estimator is asymptotically unbiased and efficient, and it is shown to converge to the true pdf under several measures. Moreover, it is a continuous and differentiable function of the elements of the unmixing matrix W , with its gradient being given by

$$\nabla p(y_i) = \frac{1}{Mh^2} \sum_{m=1}^M \mathbf{x}^{(m)} (y_i - \mathbf{w}_i \mathbf{x}^{(m)}) \phi \left(\frac{y_i - \mathbf{w}_i \mathbf{x}^{(m)}}{h} \right). \quad (8)$$

Using the kernel expansion of the source distributions, we can derive a closed form expression for the pdf estimate of the one-dimensional (1-D) reconstructed signals, evaluated at the data points as

$$p_{y_i}(\mathbf{w}_i \mathbf{x}^{(k)}) = \frac{1}{Mh} \sum_{m=1}^M \phi \left(\frac{\mathbf{w}_i (\mathbf{x}^{(k)} - \mathbf{x}^{(m)})}{h} \right). \quad (9)$$

C. Objective Function Derivation

The expectation in (4) can be approximated by its ergodic average, as follows:

$$L(W) \approx -\frac{1}{M} \sum_{i=1}^N \sum_{k=1}^M \log p_{y_i}(\mathbf{w}_i \mathbf{x}^{(k)}) - \log |\det W| \quad (10)$$

resulting in the following cost function definition

$$L(W) = -L_0(W) - \log |\det W| \quad (11)$$

where $L_0(W)$ is obtained by replacing the marginal pdfs p_{y_i} with their kernel density estimates

$$\begin{aligned} L_0(W) &= \sum_{i=1}^N E \log \left[\frac{1}{Mh} \sum_{m=1}^M \phi \left(\frac{y_i - Y_{im}}{h} \right) \right] \\ &\approx \frac{1}{M} \sum_{i=1}^N \sum_{k=1}^M \log \left[\frac{1}{Mh} \sum_{m=1}^M \phi \left(\frac{\mathbf{w}_i (\mathbf{x}^{(k)} - \mathbf{x}^{(m)})}{h} \right) \right]. \end{aligned} \quad (12)$$

The overall optimization problem can thus be posed as

$$\min_W -\frac{1}{M} \sum_{i=1}^N \sum_{k=1}^M \log \left[\frac{1}{Mh} \sum_{m=1}^M \phi \left(\frac{\mathbf{w}_i(\mathbf{x}^{(k)} - \mathbf{x}^{(m)})}{h} \right) \right] - \log |\det W| \quad (13)$$

$$\text{s.t. } \|\mathbf{w}_i\| = 1, i = 1, \dots, N. \quad (14)$$

Given the sample data $\mathbf{x}^{(k)}$, $k = 1, \dots, M$, the objective (13) is a nonlinear function of the elements of the matrix W . The additional constraints (14) are introduced in order to restrict the space of possible solutions of the problem to a finite set. Clearly, if a matrix W_0 is optimal according to (1), so is any other matrix obtained from W_0 by rescaling or permuting its rows. The constraints (14) remove the degree of freedom given by the magnitude of the sources, thus limiting the solution space to all possible permutations of the reconstructed signals (a finite set).

Although it is not strictly required in the proposed algorithm, we can assume that the mixture data has been centered and sphered prior to attempting the reconstruction [30], thus the problem is reduced to the estimation of an orthogonal matrix [10]. Such preprocessing of the mixture data allows a further simplification in the design of the kernel density estimator, since all the reconstructed signals can be assumed to be zero-mean and unit variance random variables, due to the constraint (14). Therefore, the optimal value of the parameter h , which controls the smoothness of the functional, is uniquely a function of the sample size ($h = 1.06 M^{-1/5}$, [29]). Simulation experiments reported in Section IV show a relative insensitivity of the algorithm's performance for variations up to $\pm 50\%$ from the optimal value of the bandwidth parameter.

III. OPTIMIZATION AND GLOBAL CONVERGENCE ISSUES

A. Optimization Algorithm

The objective (13) is a smooth nonlinear function of the elements w_{ij} of the unmixing matrix W . Its gradient can be computed from (8), as follows:

$$\begin{aligned} \nabla L(W) &= -\nabla L_0(W) - \nabla \log |\det(W)| \\ &= -\nabla L_0(W) - (W^T)^{-1}. \end{aligned} \quad (15)$$

If we define the following quantity:

$$Z_i(k, m) \triangleq \frac{\mathbf{w}_i(\mathbf{x}^{(k)} - \mathbf{x}^{(m)})}{h} = \frac{1}{h} \sum_{j=1}^N w_{ij}(X_{jk} - X_{jm}) \quad (16)$$

we can compute the components of $\nabla L_0(W)$ as

$$\frac{\partial L_0(W)}{\partial w_{ij}} = \frac{1}{M} \sum_{k=1}^M \frac{\sum_{m=1}^M \frac{\partial Z_i(k, m)}{\partial w_{ij}} \phi' [Z_i(k, m)]}{\sum_{m=1}^M \phi [Z_i(k, m)]} \quad (17)$$

$$= -\sum_{k=1}^M \frac{\sum_{m=1}^M (X_{jk} - X_{jm}) Z_i(k, m) \phi [Z_i(k, m)]}{h \cdot \sum_{m=1}^M \phi [Z_i(k, m)]} \quad (18)$$

since

$$\frac{\partial Z_i(k, m)}{\partial w_{ij}} = \frac{X_{jk} - X_{jm}}{h}. \quad (19)$$

The constraints (14) can be enforced by operating the substitution

$$\mathbf{w}_i = \frac{\tilde{\mathbf{w}}_i}{\|\tilde{\mathbf{w}}_i\|}, \quad i = 1, \dots, N. \quad (20)$$

Using the transformation (20), the matrix W can be written as $W = \tilde{D}^{-1} \tilde{W}$, with

$$\tilde{D} = \begin{bmatrix} \|\tilde{\mathbf{w}}_1\| & & 0 \\ & \ddots & \\ 0 & & \|\tilde{\mathbf{w}}_N\| \end{bmatrix} \quad (21)$$

thus $\tilde{W} = \tilde{D}W$. Then

$$\log |\det W| = -\sum_{i=1}^N \log \|\tilde{\mathbf{w}}_i\| + \log |\det \tilde{W}|. \quad (22)$$

The derivatives with respect to \tilde{w}_{ij} are thus computed as

$$\frac{\partial (\log |\det W|)}{\partial \tilde{w}_{ij}} = -\frac{\tilde{w}_{ij}}{\|\tilde{\mathbf{w}}_i\|^2} + [(\tilde{W}^T)^{-1}]_{ij}. \quad (23)$$

When W is orthogonal ($W^{-1} = W^T$), we have

$$(\tilde{W}^T)^{-1} = \tilde{D}^{-1}(W^T)^{-1} = \tilde{D}^{-2} \tilde{W} \quad (24)$$

and the coefficients of the gradient (23) are all equal to zero. Therefore, as expected, the second term of the cost function (3) will no longer enter the optimization procedure when the matrix W is orthogonal. Applying the substitution as in (20), the components of $\nabla L_0(\tilde{W})$ can be computed as

$$\frac{\partial L_0(\tilde{W})}{\partial \tilde{w}_{ij}} = \frac{1}{M} \sum_{k=1}^M \frac{\sum_{m=1}^M \frac{\partial \tilde{Z}_i(k, m)}{\partial \tilde{w}_{ij}} \phi' [\tilde{Z}_i(k, m)]}{\sum_{m=1}^M \phi [\tilde{Z}_i(k, m)]} \quad (25)$$

where

$$\frac{\partial \tilde{Z}_i(k, m)}{\partial \tilde{w}_{ij}} = \frac{1}{h} (X_{jk} - X_{jm} - \tilde{Z}_i(k, m) \tilde{w}_{ij}) \quad (26)$$

$$\phi' [\tilde{Z}_i(k, m)] = -\tilde{Z}_i(k, m) \phi [\tilde{Z}_i(k, m)] \quad (27)$$

TABLE I
MAIN STEPS OF THE NONPARAMETRIC ICA ALGORITHM

NON-PARAMETRIC ICA	
Initialize W, α, β	
Initialize the Hessian estimate $H := I$	
repeat	
1. Compute the search direction: $V := -H^{-1}\nabla L(W)$	
2. Backtracking: compute the step size	
$\mu := 1$	
while $L(W + \mu V) > L(W) + \alpha\mu\nabla L(W)^T V$	
$\mu := \beta\mu$	
3. Update H^{-1}	
4. Update W : $W := W + \mu V$	
until $\sqrt{-V^T \nabla L(W)} \leq \epsilon$ (stopping criterion)	

and, analogously to (16), $\tilde{Z}_i(k, m)$ is defined as

$$\tilde{Z}_i(k, m) \triangleq \frac{\tilde{w}_i(\mathbf{x}^{(k)} - \mathbf{x}^{(m)})}{h} = \frac{1}{h} \sum_{j=1}^N \tilde{w}_{ij}(X_{jk} - X_{jm}) \quad (28)$$

and $\|\tilde{w}_i\|$ is arbitrarily chosen equal to one.

A natural choice for the optimization algorithm is the quasi-Newton (QN) method [31], [32], which provides a good compromise between fast convergence, and computational payload. A *backtracking* technique is adopted for the selection of the step size. The main steps of the proposed nonparametric ICA algorithm are shown in Table I. The backtracking routine ensures convergence to the closest minimum [33], even when the objective function is not convex.

B. Analysis of the Extrema of the Cost Function for $N = 2$ Sources

A well-known result in blind signal separation is that, given the assumption of linear and instantaneous mixing, the unmixing matrix is unique up to scaling and permutations [1]. Conventionally, the unmixing operator is estimated by minimizing a cost function derived from the mutual information measure (1). Although the global minimum of (1) is known to yield the desired source separation, no proof is available to show that such a function has no local minima. On the other hand, because of the uniqueness of the separation matrix (up to permutations and scaling), proved by Comon in [1], convergence to any solution other than the global would result in a failure to separate the source signals. As it was recently pointed out in [34] and [35], this specific issue is often overlooked in other ICA frameworks, where, instead, the main concern is whether convergence to a local minimum is obtained at all for an arbitrary initial guess [36].

The problem can be studied in detail in the case of mixtures of $N = 2$ sources. In this case, the unmixing matrix W can be parametrized as follows (including implicitly the unit norm constraints on the rows of W)

$$W = \begin{bmatrix} \cos \theta_1 & \sin \theta_1 \\ \cos \theta_2 & \sin \theta_2 \end{bmatrix}. \quad (29)$$

With a slight abuse of notation we can write the cost function as

$$L(\theta_1, \theta_2) = h(\theta_1) + h(\theta_2) - \log |\det(W)| \quad (30)$$

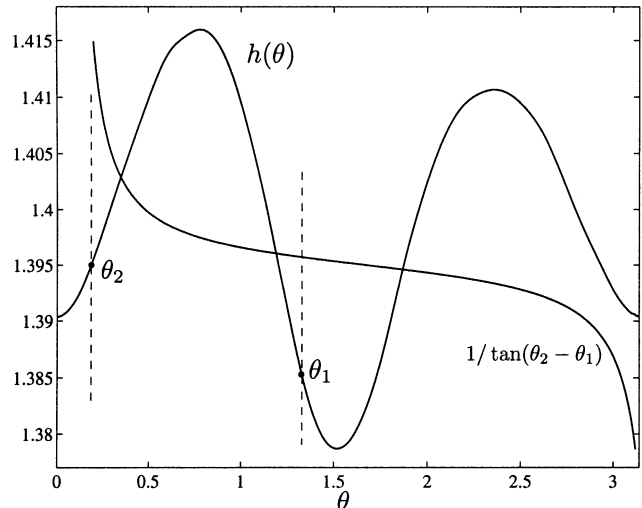


Fig. 1. Graphical interpretation of the conditions on the extrema of the cost function (34). The curve $1/\tan(\theta_2 - \theta_1)$ is plotted for a fixed value of θ_2 (not to scale).

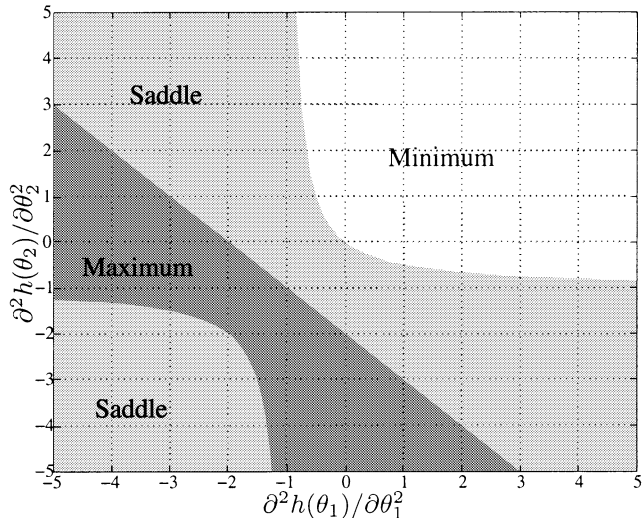


Fig. 2. The nature of an extremum of the objective function is shown as a function of the second-order partial derivatives of the entropies of the two reconstructed sources (for $\theta_2 - \theta_1 = \text{const.}$).

where $\log |\det(W)| = \log |\sin(\theta_2 - \theta_1)|$, and $h(\theta_i)$ is defined as

$$h(\theta_i) \triangleq H(y_{\theta_i}), \quad y_{\theta_i} = \cos \theta_i x_1 + \sin \theta_i x_2, \quad i = 1, 2. \quad (31)$$

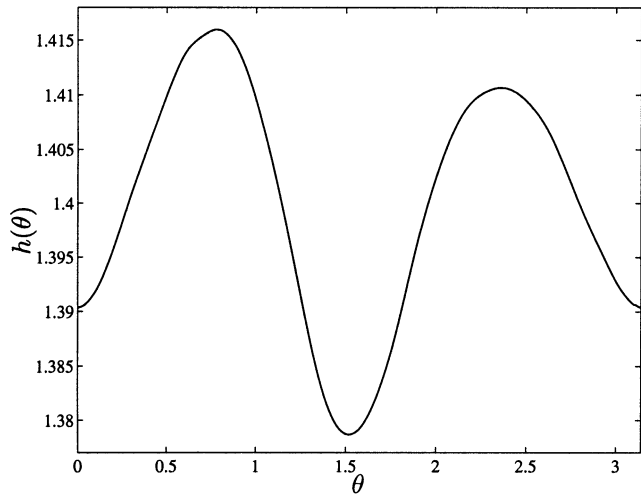
Without loss of generality, we can assume the mixing matrix to be the 2×2 identity matrix, so that $x_1 = s_1$ and $x_2 = s_2$. The extrema of cost function (30) must, then, satisfy the following conditions

$$\frac{\partial h(\theta_1)}{\partial \theta_1} + \frac{1}{\tan(\theta_2 - \theta_1)} = 0, \quad (32)$$

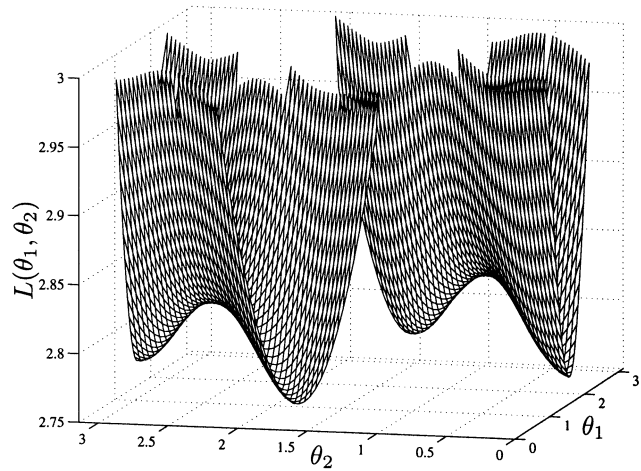
$$\frac{\partial h(\theta_2)}{\partial \theta_2} - \frac{1}{\tan(\theta_2 - \theta_1)} = 0 \quad (33)$$

or, equivalently

$$\frac{\partial h(\theta_2)}{\partial \theta_2} = -\frac{\partial h(\theta_1)}{\partial \theta_1} = \frac{1}{\tan(\theta_2 - \theta_1)}. \quad (34)$$



(a)



(b)

Fig. 3. Mixtures of a sub-Gaussian signal and a super-Gaussian signal. (a) The figure shows a plot of the entropy of a generic reconstructed source as a function of the parameter θ . For these particular mixtures of unimodal sub-Gaussian and super-Gaussian sources, the entropy function does not present any spurious local minima. (b) The overall cost function $L(W)$ is plotted as a function of (θ_1, θ_2) . The plot clearly shows the set of four equivalent minima, corresponding to permutations or change of sign of the rows of the unmixing matrix.

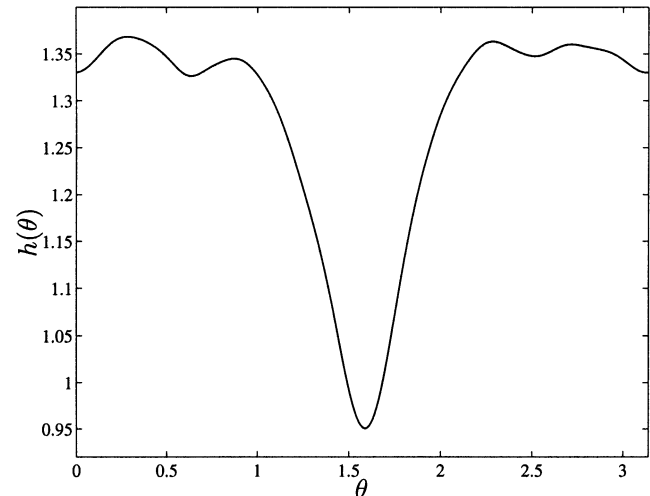
These conditions are graphically illustrated in Fig. 1. In order to characterize the nature of these extrema, we can compute the Hessian of (30), obtaining

$$\left[\frac{\partial^2 L}{\partial \theta^2} \right]_{ij} = \begin{bmatrix} \frac{\partial^2 h(\theta_1)}{\partial \theta_1^2} & 0 \\ 0 & \frac{\partial^2 h(\theta_2)}{\partial \theta_2^2} \end{bmatrix} + \frac{1}{\sin^2(\theta_2 - \theta_1)} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \quad (35)$$

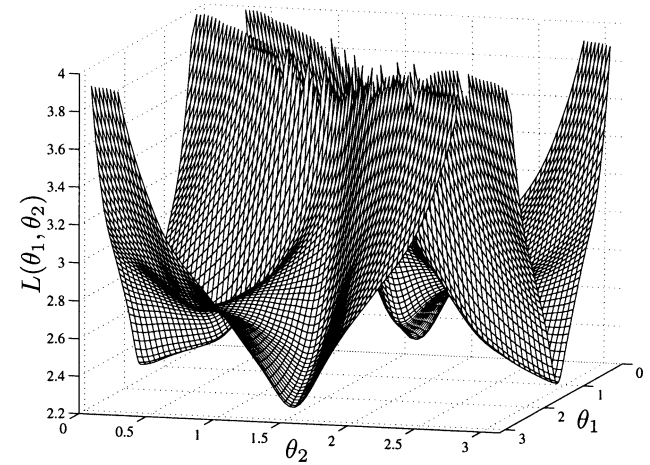
The minima of the cost function (30) are found in correspondence of values of (θ_1, θ_2) that satisfy the first-order conditions (34), and simultaneously ensure that the Hessian (35) is positive semidefinite, which requires that (see Fig. 2)

$$\frac{\partial^2 h(\theta_1)}{\partial \theta_1^2} + \frac{\partial^2 h(\theta_2)}{\partial \theta_2^2} + \frac{2}{\sin^2(\theta_2 - \theta_1)} \geq 0 \quad (36)$$

$$\frac{\partial^2 h(\theta_1)}{\partial \theta_1^2} \frac{\partial^2 h(\theta_2)}{\partial \theta_2^2} + \frac{1}{\sin^2(\theta_2 - \theta_1)} \left(\frac{\partial^2 h(\theta_1)}{\partial \theta_1^2} + \frac{\partial^2 h(\theta_2)}{\partial \theta_2^2} \right) \geq 0. \quad (37)$$



(a)



(b)

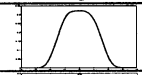
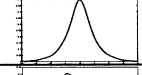
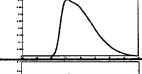
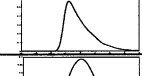


Fig. 4. Mixtures of two natural images. (a) The entropy function is plotted as a function of $\theta_{1,2}$. In this case, the entropy presents several spurious local minima, which do not correspond to independent sources. Attempting the separation using a deflationary approach could result in a failure to reconstruct the original sources. (b) The overall cost function $L(W)$ is plotted as a function of (θ_1, θ_2) , with a set of four equivalent global minima clearly appearing. The objective function is free from the spurious local minima encountered when observing the entropy function alone. At least in this case, the only values of $\theta_{1,2}$ that satisfy (34) are either (equivalent) global minima, or saddle points.

It can be easily verified that the cost function (30) is even and periodic both in θ_1 and θ_2 with period 2π , and that the conditions (34) through (37) are satisfied, in particular, when $\theta_1 = n\pi/2$ ($n \in \mathcal{Z}$), $\theta_2 = \theta_1 \pm \pi/2$, resulting in the source separation.

As an example, consider the mixture of a super-Gaussian ($\kappa_4 = 1.0$) and a sub-Gaussian source ($\kappa_4 = -1.0$), both unimodal. The entropy of an arbitrary linear projection of the mixtures is shown in Fig. 3(a) as a function of $\theta_{1,2}$ (the function is symmetric with respect to the vertical axis). Clearly, in this simple example the entropy function has only minima corresponding to the optimal solutions ($\theta_{1,2} = 0, \pm\pi/2$), which satisfy conditions (34) and (37). Because of the independence of the sources, the minima appear spaced by $\pi/2$, and correspond to the global optima of the overall cost function [see Fig. 3(b)].

The situation is quite different in the case of mixtures of sources characterized by a multimodal probability density

TABLE II
DISTRIBUTION OF THE SYNTHETIC SOURCES USED IN THE FIRST SIMULATION EXPERIMENT (SEE [38] FOR A DESCRIPTION OF THE DISTRIBUTIONS GENERATED WITH THE POWER METHOD)

Source#	Source type	Skewness	Kurtosis	Pdf plot
1	Power Exponential Distribution ($\alpha= 2.0$)	0.0	-0.8	
2	Power Exponential Distribution ($\alpha= 0.6$)	0.0	2.2	
3	Power Method Distribution ($b = 1.112, c = 0.174, d = -0.050$)	0.75	0.0	
4	Power Method Distribution ($b = 0.936, c = 0.268, d = -0.004$)	1.50	3.0	
5	Normal Distribution	0.0	0.0	
6	Rayleigh Distribution ($\beta= 1$)	0.631	0.245	

function. An interesting example is given by mixtures of natural images, where each pixel is considered as a sample drawn from a distribution. This type of sources, in fact, tend to have a distribution that is “heavily” multimodal. In Fig. 4(a), the entropy of a generic projection of a mixture of two images¹ is plotted as a function of θ . Although the entropy function shows minima at the optimal points ($\theta_{1,2} = 0, \pm\pi/2$), several spurious local minima appear in other locations. However, at least in this example, these minima do not satisfy the conditions in (34), and do not appear in the overall cost function, which, once again, has a unique set of equivalent global minima [cf. Fig. 4(b)]. The independence of the sources, in fact, imposes a special structure on the cost function, with the extrema of the entropy appearing in correspondence of orthogonal rows of the matrix W (a well known fact in the ICA theory). Other local spurious minima do not appear in the overall cost function because they do not satisfy the first-order constraints (34). Nevertheless, it is still an open problem to identify the class of distributions for which this property holds in general, as well as to show whether the same property applies for mixtures of $N > 2$ sources.

IV. SIMULATION EXPERIMENTS

A set of simulation experiments was conducted in order to investigate the performance of the proposed nonparametric method. The blind separation was attempted with each of the following algorithms: the Extended InfoMax ICA [7], FastIca [9], Jade [6], two so-called source adaptive methods, the Pearson model ICA [21] and the EGLD model ICA [37], Kernel-ICA [24] and the proposed approach². The algorithms were all downloaded from the web sites of the respective authors, and in the case of FastIca, all the available contrast functions were tested, both in deflationary mode (sources extracted one at the time), and in simultaneous separation mode

¹The images can be downloaded at <http://www.ee.ucla.edu/~riccardo/ICA/images>.

²The nonparametric ICA algorithm can be downloaded at <http://www.ee.ucla.edu/~riccardo/ICA/npica.tar.gz>.

(all the sources separated simultaneously). Both versions of Kernel-ICA, KCCA and KGV, were tested in all the simulations.

A. Mixtures of Sources With Various Distributions

In a first experiment, 1000 realizations of six different sources, distributed as shown in Table II, were independently generated, with sample sizes ranging between 500 and 5000, and mixed with randomly generated, full-rank (condition number ≤ 10) mixing matrices, noiselessly.

The separation performance was evaluated in terms of median signal-to-interference ratio (SIR), defined as $10 \log_{10} \left(\frac{\sum_{m=1}^M s_m^2}{\sum_{m=1}^M (\hat{s}_m - s_m)^2} \right)$ (dB), where s is the original signal and \hat{s} is the reconstructed signal. The “interfering” components of the reconstructed signal are by definition those that are due to sources other than the one we are attempting to separate. The results of this first experiment are shown in Fig. 5 and they clearly show the performance gain obtained with the nonparametric ICA algorithm. On the average, the *gauss* score function, when used in the simultaneous separation mode, resulted in the best overall performance for FastIca, and it is the only one reported for this first experiment. In general, SIR levels below the 8–10 dB threshold are indicative of a failure in obtaining the desired source separation.

Although the gain is more consistent in the case of skewed sources (Source #3,#4, and #6), the separation improvement is substantial also for conventional sub-Gaussian and super-Gaussian sources (Source#1 and #2). Although KernelICA-KGV appears to somehow match the performance of the proposed method, nonparametric ICA still retains a performance gain of over 5 dB on average. It is interesting to notice that, although the “source-adaptive” algorithms tend to outperform more conventional ICA methods in the case of nonsymmetric sources, they are often surpassed by traditional algorithms for symmetric ones.

The proposed technique delivers a consistent separation improvement for different sample sizes. In particular, the

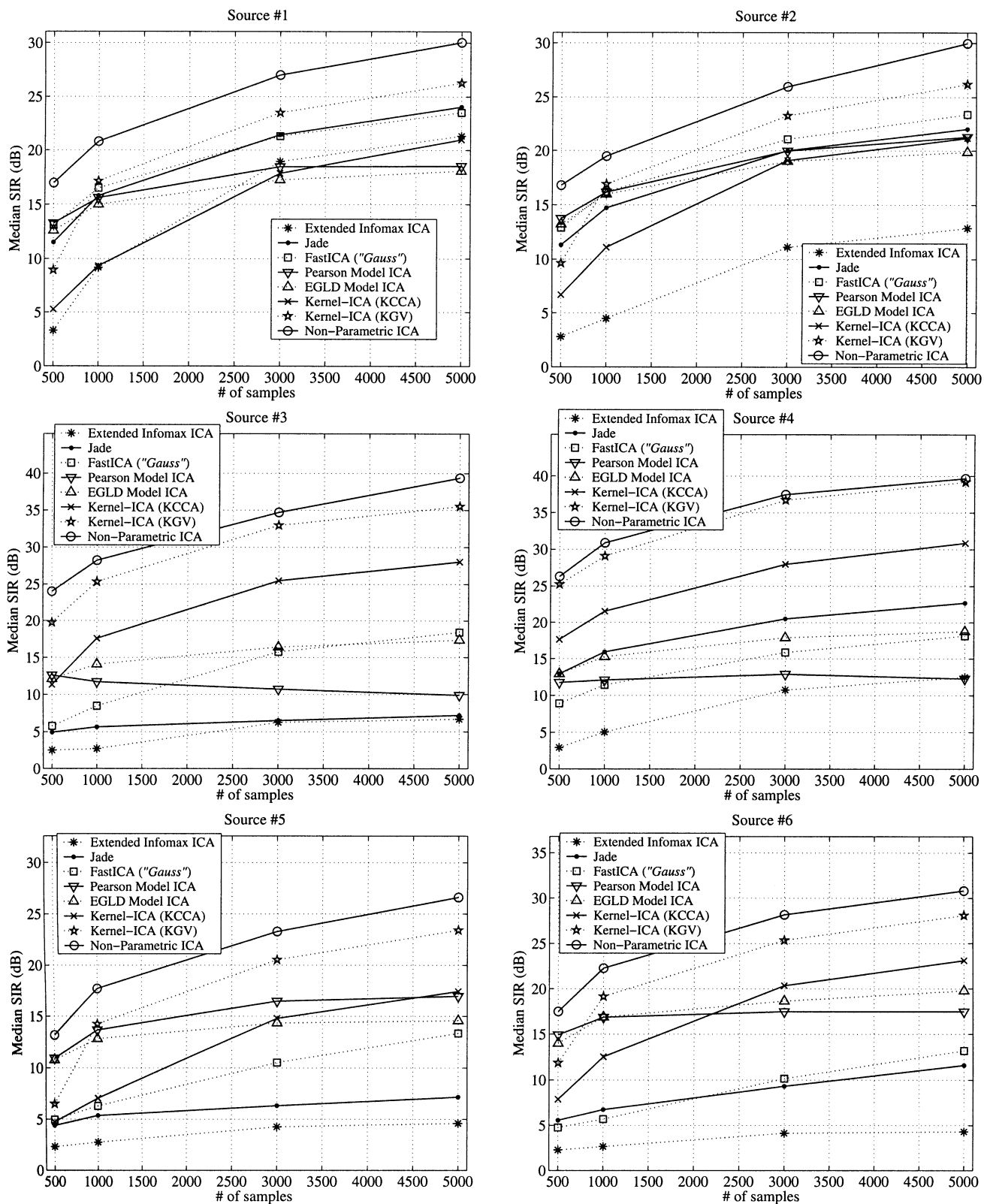


Fig. 5. First simulation experiment. The results of attempting the separation of the six different sources listed in Table II are shown for various ICA algorithms (averaged over 1000 Monte Carlo simulations). The accuracy of the separation is measured in terms of median log SIR, defined as $10 \log_{10} (\sum_{m=1}^M s_m^2 / \sum_{m=1}^M (\hat{s}_m - s_m)^2)$ (dB), where s is the original signal and \hat{s} is the reconstructed signal.

algorithm appears to be capable of learning the source statistics even when the sample size is very small (e.g., 500 samples), therefore showing promising adaptive properties.

B. Skewed Sources

In a second simulation experiment, the specific sensitivity of each algorithm to the source skewness was investigated. Using

TABLE III
THE SEPARATION PERFORMANCE IN TERMS OF MEDIAN SIR, AS WELL AS 25 AND 75 PERCENTILES, IS SHOWN FOR MIXTURES OF FOUR SKEWED SOURCES (AVERAGED OVER THE SOURCES), FOR VARIOUS ICA ALGORITHMS

ALGORITHM	SIR	25%	75%
Extended InfoMax	3.81	2.89	5.90
Jade	4.28	3.03	6.38
FastIca ('skew')	18.94	16.04	22.32
Pearson ICA	14.97	11.40	19.52
EGLD ICA	16.73	12.76	21.21
Kernel-ICA (KCCA)	16.93	13.89	20.54
Kernel-ICA (KGV)	21.64	17.86	25.10
Non-Parametric ICA	23.40	18.91	27.19

the method described in [38], we generated samples drawn from four different sources, which are characterized by a very small kurtosis ($|\kappa_4| < 0.2$), and skewness ranging between 0.0 and 0.75. The experiment was conducted mixing all four sources with randomly generated mixing matrices, using 100 independent realizations of the signals, each consisting of 2000 samples. The results obtained with the various ICA algorithms are summarized in Table III. The proposed method shows a noticeable performance improvement, confirming its capability of modeling arbitrarily distributed sources. Although FastICA resulted in the third highest median SIR, its performance is somehow biased by the choice of the score function *skew*, which assumes some *a priori* knowledge about the nature of the mixed signals.

C. Convergence Properties

The convergence properties of the algorithms were empirically tested in a third simulation experiment. The goal was to measure the approximate number of data samples required by each method to achieve a median SIR of at least 20 dB. For this purpose, we created mixtures of four independent sources with a super-Gaussian ($\kappa_4 \approx 2.2$) symmetric pdf and we averaged the separation results over 100 simulations, for different sample sizes. The choice of standard super-Gaussian sources guarantees that the experiment is unbiased, since all ICA algorithms under evaluation are capable of separating this type of signals accurately. Our results show that the proposed method is able to achieve the required quality of separation (20 dB median SIR) with only 750 samples, performance matched by KernelICA-KGV. FastICA resulted in the second-best performance (1000 samples), when the score function was suitably chosen (in this case *gauss*).

D. Bandwidth Parameter Sensitivity

The sensitivity of the algorithm to the choice of the bandwidth parameter h in (5) was evaluated following the experimental setting used in the first simulation (sources generated according to Table II). In a series of Monte Carlo simulations, the bandwidth parameter was allowed to vary up to 50% from the optimal value, computed as a function of the sample size. The results displayed in Fig. 6 show the obtained median SIR averaged across the six sources, for a sample size equal to 1000. The experiment seems to suggest that the separation performance is

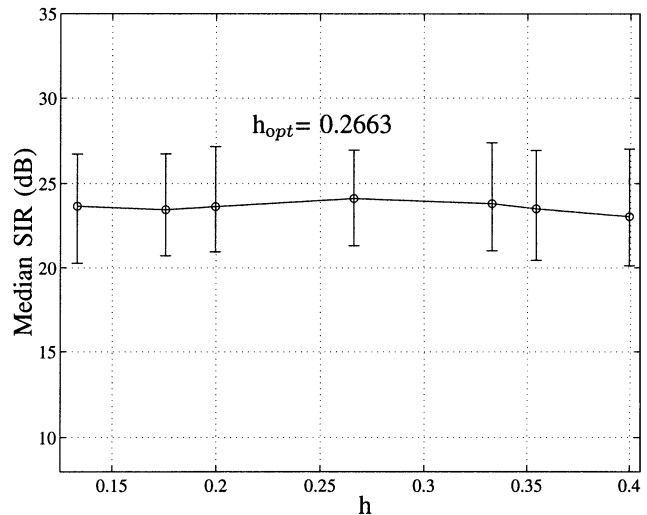


Fig. 6. The figure shows the results of a set of simulation experiments aiming at evaluating the sensitivity of the proposed technique to the choice of the bandwidth parameter h . The error bars span between the 25 and the 75 percentiles of the SIR. This experiment seems to suggest that variations of the parameter up to $\pm 50\%$ from the estimated optimal value do not considerably affect the separation performance.

relatively insensitive to the particular choice of this parameter in a broad range of values.

E. Algorithmic Complexity

The introduction of a technique enabling the simultaneous estimation of the unmixing matrix and of the unknown pdfs of the sources is inevitably accompanied by an increase in its computational complexity. Regardless of the actual optimization algorithm, a brute force implementation of the proposed nonparametric method would require an amount of floating point operations proportional to $\mathcal{O}(M^2N)$ to evaluate the cost function and $\mathcal{O}(M^2N^2)$ to compute its derivatives, where N is the number of sources and M is the sample size. This compares unfavorably with fixed score function algorithms like FastICA whose computational complexity is $\mathcal{O}(MN)$ and $\mathcal{O}(MN^2)$, respectively, especially when the number of samples M is very large.

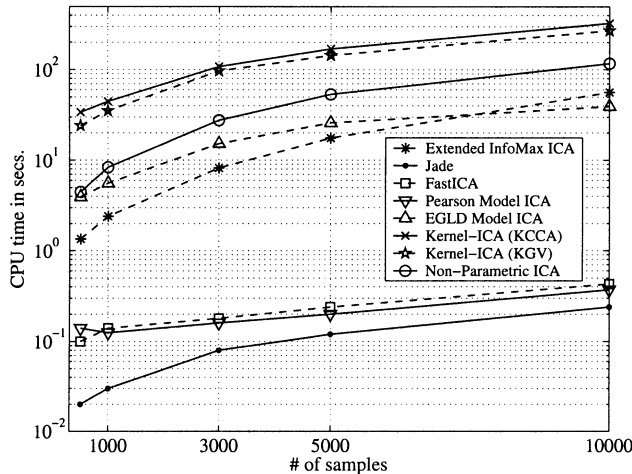
On the other hand, fast density estimation techniques based on the FFT algorithm can be developed, based on the observation that evaluating a density estimate is equivalent to computing the convolution of an unevenly sampled sequence with a Gaussian kernel [29]. At the core of the proposed nonparametric method for ICA stands a fast density estimation algorithm of this type, which can perform the evaluation of the cost function and of its derivatives in a time proportional to $\mathcal{O}(M \log_2 MN)$ and $\mathcal{O}(M \log_2 MN^2)$, respectively, thus minimizing the additional payload required to achieve the increased separation performance and reliability. Table IV shows a detailed breakdown of the computational complexity of each step of the nonparametric ICA algorithm.

The median CPU time required to run the various ICA algorithms is shown in Fig. 7(a) for a fixed number of sources ($N = 6$) and a variable number of samples and in Fig. 7(b) for a fixed number of samples ($M = 1000$) and a variable number of sources³. Clearly, fixed contrast function or moment based ICA

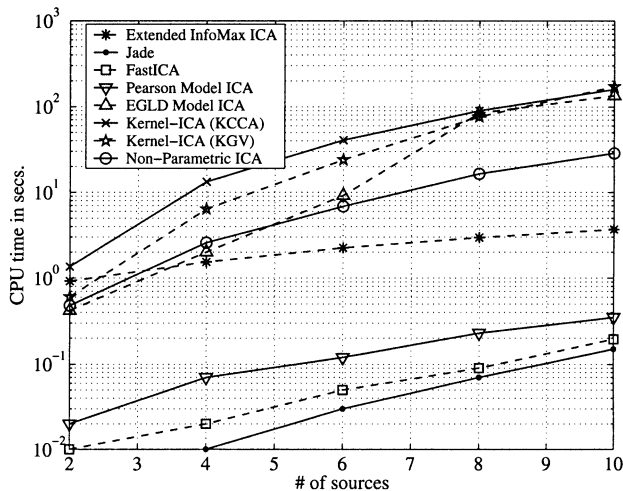
³The simulations were all performed under MATLAB; v.6.3, on a Dual Pentium IV 1.8 Ghz PC with 512 Mbytes of RAM, running Red Hat Linux v7.2.

TABLE IV
DETAILED ANALYSIS OF THE COMPUTATIONAL COMPLEXITY OF
NON-PARAMETRIC ICA AS A FUNCTION OF THE NUMBER OF SOURCES
(N) AND THE NUMBER OF SAMPLES (M)

ROUTINE	COMPLEXITY
(a) Compute search direction	$\mathcal{O}(N^4)$
(b) Backtracking routine	
Cost function evaluation (' <i>EstimateObjFFT</i> ')	
1. Data rebinning	$\mathcal{O}(NM)$
2. FFT of re-binned data	$\mathcal{O}(NM \log_2 M)$
3. FFTs multiplication	$\mathcal{O}(NM)$
4. Inverse FFT of pdf estimate	$\mathcal{O}(NM \log_2 M)$
5. Rebinning and entropy evaluation	$\mathcal{O}(NM)$
(c) Gradient computation (' <i>EstimateGradFFT</i> ')	
1. Data rebinning	$\mathcal{O}(N^2 M)$
2. FFT of re-binned data	$\mathcal{O}(N^2 M \log_2 M)$
3. FFTs multiplication	$\mathcal{O}(N^2 M)$
4. Inverse FFT of pdf derivative estimates	$\mathcal{O}(N^2 M \log_2 M)$
5. Rebinning and gradient evaluation	$\mathcal{O}(N^2 M)$
(d) Inverse Hessian update	$\mathcal{O}(N^4)$
(e) Convergence criterion evaluation	$\mathcal{O}(NM)$
Overall computational complexity:	$\mathcal{O}(N^4 + N^2 M \log_2 M)$



(a)



(b)

Fig. 7. The running time in terms of CPU seconds is shown for various ICA algorithms, for a fixed number of sources (6) and variable number of samples (a) and for a fixed number of samples (1000) and variable number of sources (b). The methods capable of source adaptation are in general computationally more expensive, as the improved separation performance is paid in terms of running time.

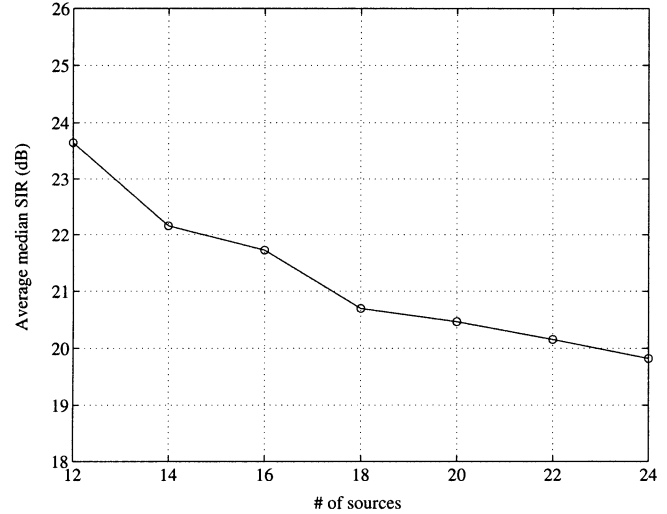


Fig. 8. Large scale simulation. The median SIR (dB) achieved by nonparametric ICA is shown for the separation of a number of sources varying between 12 and 24 (averaged over the reconstructed signals), and a fixed number of samples ($M = 1000$).

algorithms are in general significantly faster than source adaptive methods. Although nonparametric ICA is among the algorithms characterized by a higher computational complexity, it is interesting to notice that it is on average one order of magnitude faster than Kernel-ICA.

F. Large Scale Problems

In a separate simulation, we investigated the properties of the proposed method for large scale problems. This was accomplished by creating mixtures of 12 up to 24 signals, randomly chosen among a set of sources, whose distributions included both unimodal and bimodal pdfs. The separation results obtained over 100 Monte Carlo simulations (Fig. 8) demonstrate nonparametric ICA's capability of seamlessly handling large size problems. The decrease in median SIR which accompanies the increase in the problem size can be explained by considering that, while the sample size is kept constant ($M = 1000$ samples), the number of parameters that needs to be estimated ($N(N-1)/2$) increases approximately as the square of the number of sources. For example, the unmixing matrix has a total of 66 unique elements when $N = 12$, that number increasing to 276 for $N = 24$ sources.

In terms of convergence properties, we noticed only a marginal increase in the number of Newton steps required to achieve the desired separation accuracy, with the relative CPU time required to complete the routine closely matching the asymptotic computational complexity analysis described in Table IV.

V. CONCLUSION

A novel nonparametric independent component analysis algorithm was introduced. The proposed method is truly blind to the particular distribution of the original sources, and does not require the selection of optimal working parameters, or suitable nonlinearities to act as contrast functions. The algorithm outperformed state-of-the-art ICA techniques in several simulation experiments, with different types of mixtures. The capability of

modeling sources with arbitrary distribution, combined with the good convergence properties for small sample sizes, make the proposed approach a particularly attractive alternative to current ICA algorithms, especially for the analysis of real-world mixtures.

ACKNOWLEDGMENT

The authors would like to thank Prof. L. Vandenberghe for the valuable suggestions on the optimization problem.

REFERENCES

- [1] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
- [2] A. J. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
- [3] D. T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Trans. Signal Processing*, vol. 44, pp. 2768–2779, Nov. 1996.
- [4] J. P. Nadal and N. Parga, "Non-linear neurons in the low-noise limit: A factorial code maximizes information transfer," *Network*, vol. 4, pp. 295–312, 1994.
- [5] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE. Special Issue on Blind Identification and Estimation*, vol. 9, pp. 2009–2025, Oct. 1998.
- [6] —, "High-order contrasts for independent component analysis," *Neural Computat.*, vol. 11, no. 1, pp. 157–192, Jan. 1999.
- [7] T.-W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources," *Neural Computat.*, vol. 11, no. 2, pp. 417–441, 1999.
- [8] M. Girolami and C. Fyfe, "An extended exploratory projection pursuit network with linear and nonlinear anti-Hebbian lateral connections applied to the cocktail party problem," *Neural Networks*, vol. 10, no. 9, pp. 1607–1618, 1997.
- [9] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.
- [10] D. Obradovic and G. Deco, "Information maximization and independent component analysis: Is there a difference?," *Neural Computat.*, vol. 10, pp. 2085–2101, 1998.
- [11] S. Cruces, A. Cichocki, and L. Castedo, "An iterative inversion approach to blind source separation," *IEEE Trans. Neural Networks*, vol. 11, pp. 1423–1437, Nov. 2000.
- [12] R. Boscolo, H. Pan, and V. P. Roychowdhury, "Non-parametric ICA," in *Proc. Third Int. Conf. Independent Component Analysis and Blind Signal Separation*, T.-W. Lee, T.-W. Jung, S. Makeig, and T. Sejnowski, Eds., San Diego, CA, Dec. 2001, pp. 13–18.
- [13] T.-W. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real-world signals," *IEEE Proc. ICNN*, pp. 2129–2135, 1997.
- [14] A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Res.*, vol. 37, pp. 3327–3338, 1997.
- [15] M. J. McKeown, T.-P. Jung, S. Makeig, G. Brown, S. S. Kindermann, T.-W. Lee, and T. J. Sejnowski, "Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task," in *Proc. National Acad. Sciences USA*, vol. 95, 1998, pp. 803–810.
- [16] M. J. McKeown, S. Makeig, G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, V. Iragui, and T. J. Sejnowski, "Analysis of fMRI by blind separation into independent spatial components," *Human Brain Mapping*, vol. 6, no. 3, pp. 160–188, 1998.
- [17] S. Makeig, T.-P. Jung, A. J. Bell, D. Ghahremani, and T. J. Sejnowski, "Blind separation of event-related brain responses into independent components," in *Proc. National Acad. Sciences USA*, vol. 94, 1997, pp. 10979–10984.
- [18] T.-P. Jung, C. Humphries, T.-W. Lee, M. J. McKeown, V. Iragui, S. Makeig, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiol.*, vol. 37, pp. 163–178, 2000.
- [19] S.-i. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, 1997.
- [20] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEEE Proc.-F*, vol. 140, pp. 362–370, Dec. 1993.
- [21] J. Karvanen, J. Eriksson, and V. Koivunen, "Pearson system based method for blind separation," in *Proc. Second Int. Workshop on Independent Component Analysis and Blind Signal Separation*, P. Pajunen and J. Karhunen, Eds., Helsinki, 2000, pp. 585–590.
- [22] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.
- [23] N. Vlassis and Y. Motomura, "Efficient source adaptivity in independent component analysis," *IEEE Trans. Neural Networks*, vol. 12, pp. 559–566, May 2001.
- [24] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Machine Learning Res.*, vol. 3, pp. 1–48, 2002.
- [25] H. B. Barlow, "Unsupervised learning," *Neural Computation*, vol. 1, pp. 295–311, 1989.
- [26] D. Donoho, "On minimum entropy deconvolution," in *Proc. Second Applied Time Series Symp.*, D. F. Findley, Ed., New York, 1981, pp. 565–608.
- [27] J.-F. Cardoso, "Infomax and maximum likelihood for source separation," *IEEE Signal Processing Lett.*, vol. 4, pp. 112–114, Apr. 1997.
- [28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [29] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall, 1985.
- [30] M. C. Jones, "The projection pursuit algorithm for exploratory data analysis," Ph.D., School of Mathematics, Univ. Bath, 1983.
- [31] S. Boyd and L. Vandenberghe. (2000) Convex Optimization. [Online]. Available: <http://www.ee.ucla.edu/ee236b/reader.pdf>
- [32] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1999.
- [33] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer-Verlag, 1999.
- [34] M. Rattray and G. Basalyga, "Scaling laws and local minima in hebbian ica," in *Advances in Neural Information Processing Systems*, vol. 14, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., Vancouver, Canada, Dec. 2001.
- [35] —, "Stochastic trapping in a solvable model of on-line independent component analysis," *Neural Computation*, vol. 14, pp. 421–435, 2002.
- [36] S.-i. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind source separation," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1996, vol. 8, pp. 757–763.
- [37] J. Eriksson, J. Karvanen, and V. Koivunen, "Source distribution adaptive maximum likelihood estimation of ICA model," in *Proc. Second Int. Workshop on Independent Component Analysis and Blind Signal Separation*, P. Pajunen and J. Karhunen, Eds., Helsinki, 2000, pp. 227–232.
- [38] A. I. Fleishman, "A method for simulating nonnormal distributions," *Psychometrika*, vol. 43, no. 4, pp. 521–532, Dec. 1978.



Riccardo Boscolo received the "Laurea" degree in electrical engineering from the University of Padova, Italy, in 1997 (*summa cum laude*), and the M.S. and Ph.D. degrees also in electrical engineering, from the University of California, Los Angeles, in 1999 and 2003, respectively.

Between February 1999 and December 2000, he joined the research staff at the HRL Laboratories, participating in several projects involving the application of computer vision techniques to automotive applications. He is coauthor of several patents (pending) on smart airbag systems and precrash sensing systems. He is currently an Assistant Research Engineer with the University of California investigating methods for learning patterns of gene interactions and for reconstructing transcriptional regulatory networks dynamics. His research interests include computer vision, tomographic image reconstruction systems, knowledge-based image understanding, statistical learning, blind signal separation, and computational biology.



Hong Pan (S'93–M'99) received the Ph.D. degree in Electrical and Computer Engineering from Purdue University, West Lafayette, IN, in 1999.

He then joined Cornell University Medical College, Ithaca, NY, as an Assistant Professor. His research interests include mathematical statistics, biomedical image processing and analysis, and computational neuroscience. He is a co-investigator in several research programs, including the Center for Neural System of Fear and Anxiety sponsored by NIMH, Functional Neuroimaging Research Program sponsored by the DeWitt-Wallace Fund, Borderline Personality Disorder Research Program sponsored by the Swiss Foundation for Personality Disorder Research, and fMRI Localization of Psychotic Symptoms in Schizophrenia sponsored by NIMH.

Vwani P. Roychowdhury received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1989.

From 1991 to 1996, he was a faculty member with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, where he was promoted to Associate Professor in 1995. In 1996, he joined the University of California, Los Angeles, where he is currently a Professor of electrical engineering. He also serves on the faculty of the Biomedical Engineering Interdepartmental Program. His research interests include models of computation, quantum and nanoelectronic computation, quantum information processing, fault-tolerant computation, combinatorics and information theory, advanced statistical processing, and adaptive algorithms. He holds the patent for the methods and apparatus for enhancing gray scale. He has coauthored several books including *Discrete Neural Computation: A Theoretical Foundation* (Englewood Cliffs, NJ: Prentice-Hall, 1995) and *Theoretical Advances in Neural Computation and Learning* (Norwell, MA: Kluwer, 1994).

Prof. Roychowdhury was a General Motors Faculty Fellow at Purdue University from 1992 until 1994 and was awarded the Ruth and Joel Spira Outstanding Teacher Award in 1994. He received the 1999 Best Paper Award from the IEEE TRANSACTIONS ON NEURAL NETWORKS for his paper, "On relative convergence properties of principal component analysis algorithms."