Source Separation in Post-Nonlinear Mixtures

Anisse Taleb and Christian Jutten

***Abstract*—In this paper, we address the problem of separation of mutually independent sources in nonlinear mixtures. First, we propose theoretical results and prove that in the general case, it is not possible to separate the sources without nonlinear distortion. Therefore, we focus our work on specific nonlinear mixtures known as post-nonlinear mixtures. These mixtures constituted by a linear instantaneous mixture (linear memoryless channel) followed by an unknown and invertible memoryless nonlinear distortion, are realistic models in many situations and emphasize interesting properties i.e., in such nonlinear mixtures, sources can be estimated with the same indeterminacies as in instantaneous linear mixtures. The separation structure of nonlinear mixtures is a two-stage system, namely, a nonlinear stage followed by a linear stage, the parameters of which are updated to minimize an output independence criterion expressed as a mutual information criterion. The minimization of this criterion requires knowledge or estimation of source densities or of their log-derivatives. A first algorithm based on a Gram–Charlier expansion of densities is proposed. Unfortunately, it fails for hard nonlinear mixtures. A second algorithm based on an adaptive estimation of the log-derivative of densities leads to very good performance, even with hard nonlinearities. Experiments are proposed to illustrate these results.**

***Index Terms*—Entropy, neural networks, nonlinear mixtures, source separation, unsupervised adaptive algorithms.**

I. Introduction

The problem of source separation consists of retrieving unobserved sources $s_i(t), i = 1, \cdots, n$ assumed to be statistically independent from only p observed signals $x_j(t), j = 1, \cdots, p$, which are unknown functions of the sources. Contrary to other communication problems such as equalization and deconvolution, in the problem of source separation, samples of each source are not assumed to be independent and identically distributed (i.i.d.). In fact, the source distribution is unknown, and no assumption concerning the temporal dependence between samples is made or used. For this reason, the problem is generally called *blind* source separation.

The source separation problem has been intensively studied over the last 12 years for linear instantaneous (memoryless) mixtures [2], [5], [12], [16], [18], [20], [22], [25], [29] and, more recently, since 1990, for linear convolutive mixtures [13], [15], [19], [24], [36], [37], [42]. Up to now, more realistic models, especially nonlinear models, have been sketched by only a few authors. Burel [4] has proposed a neural-network based solution for the case of known nonlinearity depending on unknown parameters. Pajunen *et al.* [26] have addressed the problem using self-organizing maps. This approach, although simple and attractive, requires a huge number of neurons for good accuracy and is restricted to sources having probability density functions (pdf's) with bounded supports. Deco and Brauer [11] have also addressed the problem, considering a volume conservation condition on the nonlinear transforms. This constraint leads to very restrictive transforms and will not be considered in this paper.

More recently, Yang *et al.* [40] proposed algorithms, without separability results, for special nonlinear mixtures that were similar to post-nonlinear mixtures proposed herein in which the nonlinearity is not componentwise and whose inverse can be approximated by a two-layer perceptron.

In this paper, we investigate nonlinear mixtures, and we develop algorithms for particular (although realistic) models called post-nonlinear (PNL) models. We restrict the study to the case $n = p$, where the number of sources is then equal to the number of sensors.

In Section II, we consider the general nonlinear model of mixtures and study separability using only the statistical independence of sources. In Section III, the PNL model and its separation architecture are presented. The derivation of independence criteria, assuming the knowledge of source distributions, is developed in Section IV. Section V addresses estimation of pdf and score functions. Combining these results leads to practical algorithms and experiments described in Section VI. A discussion and remarks on future work finishes the paper.

II. Nonlinear Mixtures

Consider an n-sensor array that provides the signal $\boldsymbol{e}(t) = (e_1(t), e_2(t), \cdots, e_n(t))^T$. The signal $\boldsymbol{e}(t)$ is a nonlinear memoryless mixture of n unknown statistically independent sources $\boldsymbol{s}(t) = (s_1(t), s_2(t), \cdots, s_n(t))^T$ if it can be written as

$$e_i(t) = \mathcal{F}_i(s_1(t), \cdots, s_n(t)), \qquad i = 1, \cdots, n \tag{1}$$

where we assume that the mapping $\boldsymbol{\mathcal{F}}$ consisting of the functions $\mathcal{F}_i$ is an unknown differentiable bijective mapping from a subset of $\mathbb{R}^n$ in a subset of $\mathbb{R}^n$. For the sake of simplicity, because the mixing system is memoryless and the temporal dependence of the samples is not used, the time dependence of the variables will be omitted, and the model will be written

$$\boldsymbol{e} = \boldsymbol{\mathcal{F}}(\boldsymbol{s}). \tag{2}$$

Manuscript received January 6, 1998; revised March 12, 1999. The associate editor coordinating the review of this paper and approving it for publication was Prof. Arnab K. Shaw.

The authors are with LIS-INPG, Grenoble, France. (e-mail: taleb@lis-viallet.inpg.fr; chris@lis-viallet.inpg.fr).

Publisher Item Identifier S 1053-587X(99)07662-X.

The first question of interest concerns separability. Is it possible, using only the statistical independence assumption, to recover the sources $\boldsymbol{s}$ from the nonlinear mixture (2)?

A. Results for Linear Mixtures

For a linear memoryless mixture, (2) becomes

$$\boldsymbol{e} = \boldsymbol{A}\boldsymbol{s} \tag{3}$$

where $\boldsymbol{A}$ is a square nonsingular matrix. Source separation consists then of estimating a square nonsingular matrix $\boldsymbol{B}$ such that

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{e} = \boldsymbol{B}\boldsymbol{A}\boldsymbol{s} \tag{4}$$

has statistically independent components. $\boldsymbol{B}$ is called a separating matrix. Denoting $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}$, it is well known [8] that using an independence criterion, we can only estimate sources, when at most one is Gaussian, up to any diagonal matrix $\boldsymbol{D}$ (scale factor) and up to any permutation matrix $\boldsymbol{P}$. In other words, separation is achieved when $\boldsymbol{C} = \boldsymbol{P}\boldsymbol{D}$. This property is a direct consequence of the Darmois–Skitovich theorem [10], [31].

B. Nonlinear Mixtures

In this case, statistical independence is not a strong enough assumption to recover the sources without distortion. In fact, when X and Y are two independent random variables, (i.e., $p_{XY}(u, v) = p_X(u)p_Y(v)$) and f and g are two bijective derivable functions, then

$$p_{f(X), g(Y)}(u, v) = \frac{p_X(f^{-1}(u))}{(f' \circ f^{-1})(u)} \frac{p_Y(g^{-1}(v))}{(g' \circ g^{-1})(v)}. \tag{5}$$

This shows clearly that the random variables $f(X)$ and $g(Y)$ are also independent, and consequently, sources can be recovered only up to any nonlinear function. Such indeterminacy may lead to strong nonlinear distortions on the estimated sources and is not acceptable.

Now, if we consider again (1) and a separating mapping $\boldsymbol{\mathcal{G}}$, providing a random vector $\boldsymbol{y}$ with independent components

$$\boldsymbol{y} = \boldsymbol{\mathcal{G}}(\boldsymbol{e}) = (\boldsymbol{\mathcal{G}} \circ \boldsymbol{\mathcal{F}})(\boldsymbol{s}) = \boldsymbol{\mathcal{H}}(\boldsymbol{s}) \tag{6}$$

the question of interest is whether all mappings $\boldsymbol{\mathcal{H}}$, transforming $\boldsymbol{s}$ into $\boldsymbol{y}$ with independent components, have a diagonal Jacobian.

The answer is negative because we can construct, in an infinite number of manners, transforms with nondiagonal Jacobian conserving statistical independence [9]. Such mappings satisfy the following differential functional equation:

$$\begin{aligned} p_S(\boldsymbol{s}) &= \prod_{i=1}^{n} p_{S_i}(s_i) \\ &= |\det \boldsymbol{J}_{\mathcal{H}}| \prod_{i=1}^{n} p_{Y_i}(h_i(s_1, s_2, \cdots, s_n)). \end{aligned} \tag{7}$$

This result is well known for linear mixtures, where it is restricted to the Gaussian distribution. In that case, if $\boldsymbol{s}$ is Gaussian (the s_i components are Gaussian independent) with a covariance matrix $E[\boldsymbol{s}\boldsymbol{s}^T] = \boldsymbol{I}$ (where $\boldsymbol{I}$ is the identity matrix), then any orthogonal matrix $\boldsymbol{U}$ applied to $\boldsymbol{s}$ will conserve the statistical independence since $E[\boldsymbol{y}\boldsymbol{y}^T] = \boldsymbol{U}E[\boldsymbol{s}\boldsymbol{s}^T]\boldsymbol{U}^T = \boldsymbol{I}$.

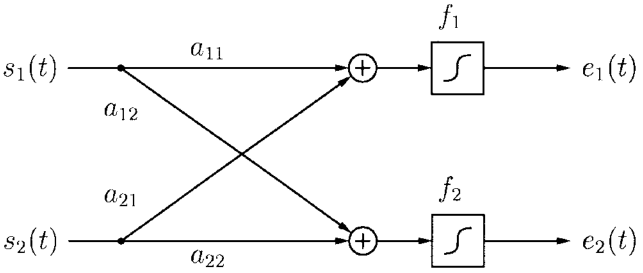


Fig. 1. Post-nonlinear mixing system ($n = 2$).

In the nonlinear case, for the sake of simplicity, we give only a simple example in the two-dimensional (2-D) case. Consider two independent Gaussian scalar random variables X and Y, whose joint pdf is

$$p_{XY}(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right), \qquad (x, y) \in \mathbb{R}^2 \tag{8}$$

and consider the nonlinear transform

$$\begin{cases} X = r \cos \theta \\ Y = r \sin \theta \end{cases} \tag{9}$$

with $r \in \mathbb{R}^+$ and $\theta \in [0, 2\pi[$. This transform has a full-rank Jacobian matrix provided that $r \neq 0$

$$\boldsymbol{J} = \begin{bmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{bmatrix} \tag{10}$$

and the joint pdf of R and Θ is then

$$p_{R, \Theta}(r, \theta) = \begin{cases} \dfrac{r}{2\pi} e^{-r^2}, & (r, \theta) \in \mathbb{R}^+ \times [0, 2\pi[\\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Relation (11) shows that the two random variables R and Θ are statistically independent. Other examples can be found in the literature (see for example Lukacs [23]) or can be easily constructed.

C. Conclusion

Source separation in the nonlinear case is, in general, impossible. The source independence assumption, which is sufficient to restore a copy of the sources in the linear case, is not strong enough in the general nonlinear case. From Darmois's results [9], we can say that there exists an infinity of mappings with *a priori* nondiagonal Jacobian matrices preserving the independence property.

III. Post-Nonlinear Mixtures

In this paper, we address particular nonlinear mixtures, which can be considered to be a hybrid model consisting of a linear stage followed by a nonlinear stage (see Fig. 1).

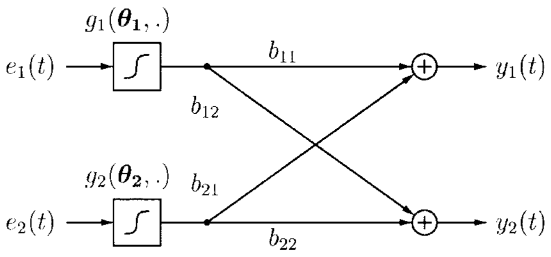


Fig. 2. Separation structure.

A. The PNL Model

This model, which was introduced by Taleb and Jutten [34], provides the observation $\boldsymbol{e}(t) = (e_1(t), e_2(t), \cdots, e_n(t))^T$, which is the unknown nonlinear mixture of the unknown statistically independent source $\boldsymbol{s}(t) = (s_1(t), s_2(t), \cdots, s_n(t))^T$

$$e_i(t) = f_i\left(\sum_{j=1}^{n} a_{ij} s_j(t)\right), \qquad i = 1, \cdots, n \tag{12}$$

where f_i are unknown invertible derivable nonlinear functions, and a_{ij} $(i, j = 1, \cdots, n)$ denote the scalar entries of a regular mixing matrix $\boldsymbol{A}$. In the following, the mixture vector $\boldsymbol{e}(t)$, and by extension the pair $(\boldsymbol{A}, \boldsymbol{f})$, will be called a post-nonlinear mixture (PNL). Although particular, this model is realistic enough. It corresponds to systems in which the transmission across a channel is linear and memoryless (this provides linear memoryless mixtures of sources), whereas sensors and their preamplifiers introduce memoryless nonlinear distortions. For instance, systems with memoryless nonlinearities can be encountered in sensor arrays [27], in digital satellite and microwave communications [30], and in some biological models [21].

B. Separability

Contrary to general nonlinear mixtures, the PNL mixtures have a favorable separability property. In fact, using the separation structure $(\boldsymbol{g}, \boldsymbol{B})$ shown in Fig. 2, it can be demonstrated, under weak conditions on the mixing matrix $\boldsymbol{A}$ and on the source distribution, that the output independence can be obtained if and only if $\forall i = 1, 2, \cdots, n$, $f_i \circ g_i$ are linear. This means that the sources $\boldsymbol{y}$, which were estimated using an independence criterion on the outputs, are equal to the unknown sources with the same indeterminacies noted in linear memoryless mixtures

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{s} + \boldsymbol{t} \tag{13}$$

where $\boldsymbol{P}$ and $\boldsymbol{\Lambda}$ are permutation and diagonal matrices, respectively, and $\boldsymbol{t}$ is a constant translation vector.

In this subsection, we give the separability Lemma with a few comments.

Lemma 1: Let $(\boldsymbol{A}, \boldsymbol{f})$ be a PNL mixture and $(\boldsymbol{g}, \boldsymbol{B})$ a PNL separation structure, where we have the following.

- $\boldsymbol{A}$ is a regular matrix and has at least two nonzero entries per row or per column.
- $f_i, i = 1, \cdots, n$ are differentiable invertible functions.
- $\boldsymbol{B}$ is a regular matrix.
- $h_i = g_i \circ f_i$ satisfies $\forall u \in \mathbb{R}$, $h_i'(u) \neq 0$, for all $i = 1, \cdots, n$.

Suppose that each source s_i accepts a density function that vanishes at one point at least, and then, the output $\boldsymbol{y}$ of the separation structure has mutually independent components if and only if the h_i components are linear, and $\boldsymbol{B}$ is a separating matrix. ■

The proof of this Lemma is given in the Appendix.

The condition on the source distributions is not restrictive for actual signals. This is especially true for signals having a bounded or a discrete distribution.

The condition on matrix $\boldsymbol{A}$, although surprising at first glance, is easy to understand. In fact, if the mixing matrix is diagonal, the PNL observations are

$$e_i(t) = f_i(a_{ii}s_i(t)), \quad \forall i = 1, \cdots, n. \tag{14}$$

Clearly, each observation e_i is a nonlinear function of the source s_i only. Thus, the observations e_i, $\forall i = 1, \cdots, n$ are already mutually independent, and the sources s_i can be restored only up to a nonlinear distortion.

If the matrix $\boldsymbol{A}$ can be written by blocks $\left[\begin{smallmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda} \end{smallmatrix}\right]$, where $\boldsymbol{A}_1$ is a regular $n_1 \times n_1$ matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix, the previous indeterminacy holds for the sources s_i, $i = n_1+1, \cdots, n$. This fact remains true if we apply a permutation to the diagonal matrix $\boldsymbol{\Lambda}$.

IV. Derivation of the Learning Rule

A. Independence Criterion

The statistical independence of the sources is the main assumption. Then, any separation structure is tuned so that the components of its output $\boldsymbol{y}$ become statistically independent. This is achieved if the joint density factorizes as the product of the marginal densities

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} p_{Y_i}(y_i). \tag{15}$$

As proposed by a few authors [8], [28], [38], [41], statistical independence can be measured using the Kullback–Liebler (KL) divergence between $p_{\boldsymbol{Y}}$ and $\prod_{i=1}^{n} p_{Y_i}$

$$\mathrm{KL}\left(p_{\boldsymbol{Y}}, \prod_{i=1}^{n} p_{Y_i}\right) = \int p_{\boldsymbol{Y}}(\boldsymbol{u}) \log \frac{p_{\boldsymbol{Y}}(\boldsymbol{u})}{\prod_{i=1}^{n} p_{Y_i}(u_i)}\, d\boldsymbol{u} \tag{16}$$

which is equal to Shannon's mutual information $I(\boldsymbol{Y})$ between the components of output vector $\boldsymbol{y}$. This can be rewritten as

$$I(\boldsymbol{Y}) = \sum_{i=1}^{n} H(Y_i) - H(\boldsymbol{Y}) \tag{17}$$

where $H(U) = -\int p_U(u) \log p_U(u)\, du$ denotes the entropy of U.

The quantity $I(\boldsymbol{Y})$ provides a measure of independence of the components of $\boldsymbol{y}$. In fact, it is always positive and vanishes iff $p_{\boldsymbol{Y}} = \prod_{i=1}^{n} p_{Y_i}$. However, the use of mutual information as a cost function is not easy because the marginal entropies

$H(Y_i)$ depend directly on the marginal densities p_{Y_i}, which are unknown and vary when minimizing $I(\boldsymbol{Y})$. In the next section, we focus on the estimation of the separation structure for PNL mixtures, assuming that the densities p_{Y_i} are known. We will see later, in Section V, how to overcome this assumption using two different approaches.

B. Parameter Estimation

1) Preliminaries: The global separation system for PNL mixtures (Fig. 2) consists of two parts:

- a **nonlinear stage**, consisting of n parametric bijective nonlinear functions $g_i(\boldsymbol{\theta}_i, u)$, $i = 1, \cdots, n$, which should cancel the channel post-distortions f_i, $i = 1, \cdots, n$. $\boldsymbol{\theta}_i$ denotes a parameter vector that is tuned to achieve output mutual independence by minimizing $I(\boldsymbol{Y})$.
- a **linear stage**, consisting of a regular matrix $\boldsymbol{B}$ devoted to the separation of the linear mixture. The matrix $\boldsymbol{B}$ is also estimated by minimizing $I(\boldsymbol{Y})$.

With respect to this separation structure, the joint pdf of the output vector $\boldsymbol{y}$ is

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{p_{\boldsymbol{E}}(\boldsymbol{e})}{|\det(\boldsymbol{B})| \prod_{i=1}^{n} |g_i'(\boldsymbol{\theta}_i, e_i)|} \tag{18}$$

and leads to the following expression of the joint entropy:

$$H(\boldsymbol{Y}) = H(\boldsymbol{E}) + \sum_{i=1}^{n} E[\log |g_i'(\boldsymbol{\theta}_i, e_i)|] + \log |\det(\boldsymbol{B})|. \tag{19}$$

The minimization of $I(\boldsymbol{Y})$ requires the computation of its gradient with respect to the separation structure parameters $\boldsymbol{B}$ and $\boldsymbol{\theta}_i$, $i = 1, \cdots, n$. The following general lemma is affected by the derivation of marginal entropies. The proof can be found in the Appendix.

Lemma 2: Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ be a random variable, and let $y = h(\boldsymbol{\gamma}, \boldsymbol{x})$ be a function of $\boldsymbol{x}$ differentiable with respect to the nonrandom parameter $\boldsymbol{\gamma}$ and such that Y accepts a differentiable pdf p_Y; then

$$\frac{dH(Y)}{d\boldsymbol{\gamma}} = -E\left[\frac{dh(\boldsymbol{\gamma}, \boldsymbol{x})}{d\boldsymbol{\gamma}} \psi_Y(h(\boldsymbol{\gamma}, \boldsymbol{x}))\right]$$

where $\psi_Y(u) = (p_Y'(u)/p_Y(u))$ is called the score function of Y. ■

2) Linear Stage: To estimate the linear stage parameters, we must compute

$$\frac{\partial I(\boldsymbol{Y})}{\partial \boldsymbol{B}} = \sum_{i=1}^{n} \frac{\partial H(Y_i)}{\partial \boldsymbol{B}} - \frac{\partial H(\boldsymbol{Y})}{\partial \boldsymbol{B}}. \tag{20}$$

We first calculate the gradient of the second term $H(\boldsymbol{Y})$ with respect to $\boldsymbol{B}$. According to (19), and since $H(\boldsymbol{E})$ and $g_i'(\boldsymbol{\theta}_i, e_i)$ $(i = 1, \cdots, n)$ do not depend on $\boldsymbol{B}$, the second term of the right-hand side of (20) is

$$\frac{\partial H(\boldsymbol{Y})}{\partial \boldsymbol{B}} = \frac{\partial \log |\det(\boldsymbol{B})|}{\partial \boldsymbol{B}} = \boldsymbol{B}^{-T}. \tag{21}$$

The first term in (20) can be simplified as $y_j = \sum_{k=1}^{n} b_{jk} x_k$, where $x_k = g_k(\boldsymbol{\theta}_k, e_k)$, and only depends on the jth row of $\boldsymbol{B}$, i.e., on b_{jk}, $k = 1, \cdots, n$

$$\frac{\partial \sum_{i=1}^{n} H(Y_i)}{\partial b_{jk}} = \frac{\partial H(Y_j)}{\partial b_{jk}}. \tag{22}$$

Using Lemma 2 on the linear model $y_j = \sum_{k=1}^{n} b_{jk} x_k$, (22) becomes

$$\frac{\partial H(Y_j)}{\partial b_{jk}} = -E[x_k \psi_{Y_j}(y_j)]. \tag{23}$$

Denoting $\boldsymbol{\psi} = (\psi_{Y_1}(y_1), \psi_{Y_2}(y_2), \cdots, \psi_{Y_n}(y_n))^T$, (23) yields

$$\frac{\partial \sum_{j=1}^{n} H(Y_j)}{\partial \boldsymbol{B}} = -E[\boldsymbol{\psi} \boldsymbol{x}^T]. \tag{24}$$

Finally, combining (21) and (24), the gradient of $I(\boldsymbol{Y})$ with respect to $\boldsymbol{B}$ is simply

$$\frac{\partial I(\boldsymbol{Y})}{\partial \boldsymbol{B}} = -E[\boldsymbol{\psi} \boldsymbol{x}^T] - \boldsymbol{B}^{-T}. \tag{25}$$

This is the same expression as in the linear source separation.

3) Nonlinear Stage: The derivation of the mutual information (17) with respect to parameters $\boldsymbol{\theta}_k$ of the nonlinear function $g_k(\boldsymbol{\theta}_k, e_k)$ is

$$\frac{\partial I(\boldsymbol{Y})}{\partial \boldsymbol{\theta}_k} = \sum_{i=1}^{n} \frac{\partial H(Y_i)}{\partial \boldsymbol{\theta}_k} - \frac{\partial H(\boldsymbol{Y})}{\partial \boldsymbol{\theta}_k}. \tag{26}$$

Writing $y_i = \sum_{j=1}^{n} b_{ij} g_j(\boldsymbol{\theta}_j, e_j)$ and again using Lemma 2, the first term becomes

$$\sum_{i=1}^{n} \frac{\partial H(Y_i)}{\partial \boldsymbol{\theta}_k} = -\sum_{i=1}^{n} E\left[\psi_{Y_i}(y_i) \frac{\partial \sum_{j=1}^{n} b_{ij} g_j(\boldsymbol{\theta}_j, e_j)}{\partial \boldsymbol{\theta}_k}\right]. \tag{27}$$

Using (19), the second term is

$$\frac{\partial H(\boldsymbol{Y})}{\partial \boldsymbol{\theta}_k} = E\left[\frac{\partial}{\partial \boldsymbol{\theta}_k} \log |g_k'(\boldsymbol{\theta}_k, e_k)|\right]. \tag{28}$$

Combining the two terms, (26) becomes

$$\begin{aligned} \frac{\partial I(\boldsymbol{Y})}{\partial \boldsymbol{\theta}_k} = &-E\left[\frac{\partial \log |g_k'(\boldsymbol{\theta}_k, e_k)|}{\partial \boldsymbol{\theta}_k}\right] \\ &- E\left[\left(\sum_{i=1}^{n} \psi_{Y_i}(y_i) b_{ik}\right) \frac{\partial g_k(\boldsymbol{\theta}_k, e_k)}{\partial \boldsymbol{\theta}_k}\right]. \end{aligned} \tag{29}$$

Of course, the complete computation of $(\partial I(\boldsymbol{Y})/\partial \boldsymbol{\theta}_k)$ depends on the structure of the parametric nonlinear mapping $g_k(\boldsymbol{\theta}_k, e_k)$. In Section VI-A, this computation is given when the mapping is estimated by a multilayer perceptron.

V. PDF and Score Functions

The derivatives of $I(\boldsymbol{Y})$ with respect to $\boldsymbol{B}$ (25) and to $\boldsymbol{\theta}_k$ (29) point out the importance of the ψ_{Y_i} functions called *score functions*

$$\psi_{Y_i}(u) = \frac{d}{du} \log p_{Y_i}(u) = \frac{p'_{Y_i}(u)}{p_{Y_i}(u)}. \tag{30}$$

Recently, Cardoso *et al.* [6] showed that the nonlinearities minimizing the rejection rates of equivariant algorithms are proportional to ψ_S in the case of i.i.d. sources $\boldsymbol{s}$. Earlier, Pham *et al.* [29] showed for the linear instantaneous mixtures using a maximum likelihood approach that the optimal nonlinear functions for source separation algorithms are the score functions of the sources. Charkani *et al.* [7] extended these results to the convolutive mixtures.

Unfortunately, these score functions are unknown and must be estimated (adaptively) from the output vector $\boldsymbol{y}$.

A first idea is to estimate the pdf p_{Y_i}, for instance, using kernel estimators, and then by derivation, we deduce an estimate of ψ_{Y_i}. This approach has been tested, but the derivation provides a noisy estimation of ψ_{Y_i}.

Pham *et al.* [29] and Charkani *et al.* [7] proposed algorithms in which the source score functions ψ_S are estimated by a linear parametric model corresponding to the projection of these functions in a subspace spanned by a few basis nonlinear functions. The difficulty of this approach lies in the choice of the basis functions.

In this paper, two approaches are considered. The first is based on a Gram–Charlier approximation of the pdf. The second approach is based on a direct estimation of ψ_{Y_i}.

A. Approximation of ψ_{Y_i} Using a Gram–Charlier Expansion

The Gram–Charlier expansion of a pdf $p_X(x)$ consists of writing the pdf as a polynomial series expansion

$$p_X(x) = \sum_{i=0}^{\infty} a_i H_i(x) N(x) \tag{31}$$

where $N(x) = (1/\sqrt{2\pi})e^{-(1/2)x^2}$ denotes the standard normal distribution, and $H_i(x)$ is the ith Chebyshev–Hermite polynomial. These polynomials are orthogonal and are defined by

$$(-1)^i \frac{d^i}{dx^i} N(x) = H_i(x) N(x). \tag{32}$$

The a_i coefficients can be computed using the orthogonality property of the H_i polynomials (see [32]). For a zero mean random variable X

$$\begin{cases} a_0 = 1 \\ a_1 = 0 \\ a_2 = \frac{1}{2}(E[x^2] - 1) \\ a_3 = \frac{1}{6}E[x^3] \\ a_4 = \frac{1}{24}(E[x^4] - 6E[x^2] + 3) \\ \vdots \quad \vdots \end{cases} \tag{33}$$

and

$$p_X(x) = N(x)\left[1 + \tfrac{1}{2}(E[x^2] - 1)H_2(x) + \tfrac{1}{6}E[x^3]H_3(x) + \tfrac{1}{24}(E[x^4] - 6E[x^2] + 3)H_4(x) + \cdots\right]. \tag{34}$$

Assuming the terms of orders greater than 4 can be neglected, i.e., the distribution is close to Gaussian, we can write

$$\log p_X(x) = -\tfrac{1}{2}x^2 + \tfrac{1}{2}(E[x^2] - 1)H_2(x) + \tfrac{1}{6}E[x^3]H_3(x) + \tfrac{1}{24}(E[x^4] - 6E[x^2] + 3)H_4(x) \tag{35}$$

and deduce

$$\psi_x(x) = -x - (E[x^2] - 1)H_1(x) + \tfrac{1}{2}E[x^3]H_2(x) + \tfrac{1}{6}(E[x^4] - 6E[x^2] + 3)H_3(x). \tag{36}$$

If X has a standard deviation $\sigma_X = 1$, then (36) may be simplified to

$$\psi_x(x) = -x + \frac{\kappa_3}{2}(x^2 - 1) + \frac{\kappa_4}{6}(x^3 - 3x) \tag{37}$$

where κ_i is the ith-order cumulant.

This kind of expansion has already been used by Comon [8] and Gaeta *et al.* [14] in the linear case and by Taleb and Jutten [34] and Yang *et al.* [39] in the nonlinear source separation case. The main advantages of this approach are its simplicity and its low computational cost.

Although the methods based on this expansion work very well for linear source separation, this approximation leads to a poor estimate of the separation structure in the nonlinear case. More precisely, fair results may be obtained only if the unknown nonlinear functions $f_i(.)$ do not imply excessively hard distortions (see Fig. 3). Conversely, estimates are of very poor quality (see Fig. 4) if the distortion is hard. A better approximation could be obtained using terms of orders higher than 4, but the main advantage (simplicity) of the approach would be lost.

We may remark that (37) is a truncated polynomial expansion depending on 1, x, x^2, and x^3 and is thus not optimal. It could be improved to produce an optimal one (in the mean square error sense) by computing parameters ψ_0, ψ_1, ψ_2, and ψ_3 minimizing $E[(\hat{\psi}_X(x) - \psi_X(x))^2]$ with $\hat{\psi}_X(x) = \psi_0 + \psi_1 x + \psi_2 x^2 + \psi_3 x^3$ (see Section V-B2 for more detailed computations).

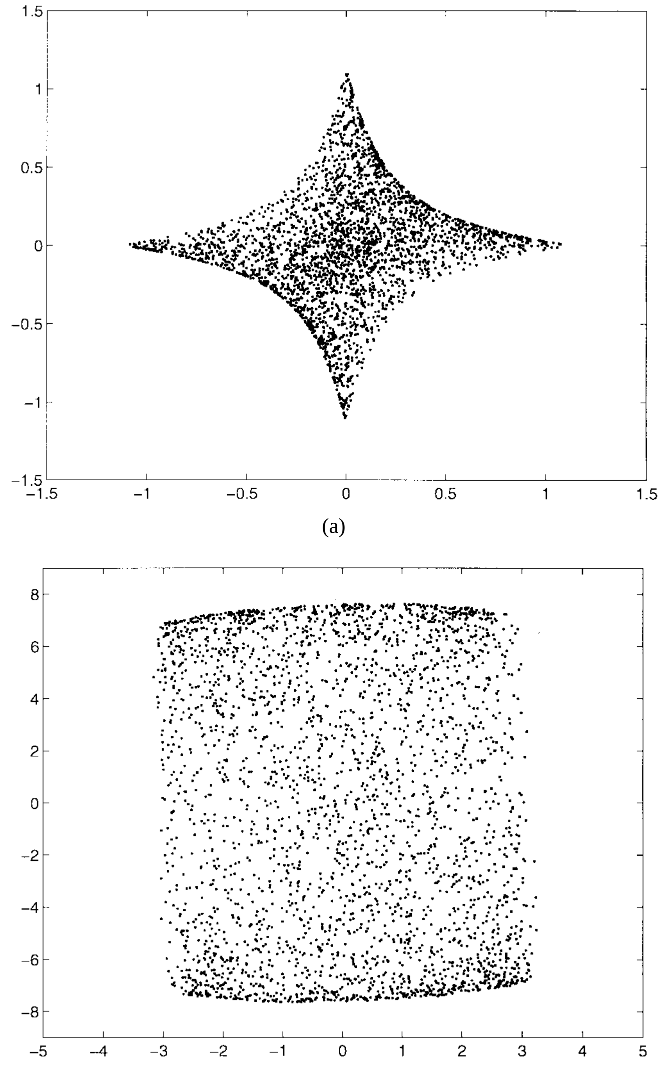
B. Direct Estimation of ψ_{Y_i}

1) LMS Estimation: We thus propose a second approach consisting of a direct estimation of ψ_Y based on a nonlinear parametric model $\phi(\boldsymbol{w}, u)$. The parameter vector $\boldsymbol{w}$ is estimated by minimizing the mean square error

$$\mathcal{E}(\boldsymbol{w}) = \tfrac{1}{2}E[(\phi(\boldsymbol{w}, y) - \psi_Y(y))^2] \tag{38}$$

according to a gradient descent algorithm

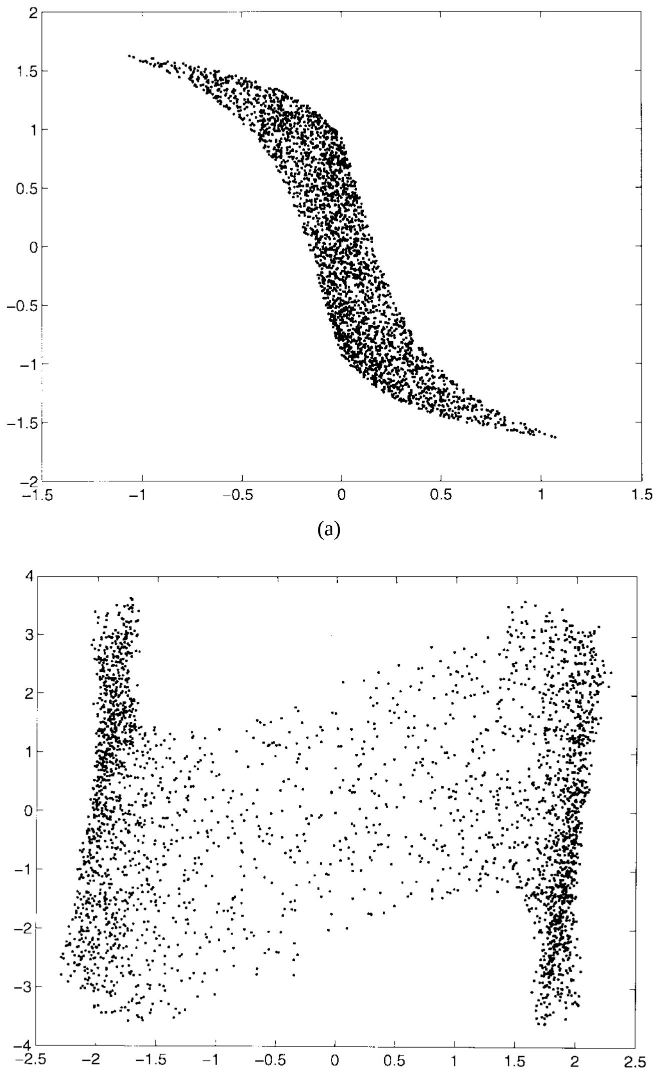
$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \mu_t \left(\frac{\partial \mathcal{E}}{\partial \boldsymbol{w}}\right)\Bigg|_{\boldsymbol{w}=\boldsymbol{w}(t)} \tag{39}$$



(a)

(b)

Fig. 3. Result with Gram–Charlier approximation obtained with soft PNL distortions. (a) Mixture distribution and (b) output distribution show output quasi-independence.



(a)

(b)

Fig. 4. Result with Gram–Charlier approximation obtained with hard distortions. (a) Mixture distribution and (b) output distribution show that output independence is not achieved.

where μ_t is a positive adaptation step. The gradient of $\mathcal{E}(\boldsymbol{w})$ with respect to the parameter vector $\boldsymbol{w}$ is expressed as

$$\begin{aligned}\frac{\partial \mathcal{E}(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{w}} E[(\phi(\boldsymbol{w}, y) - \psi_Y(y))^2] \\ &= E\left[\frac{\partial \phi}{\partial \boldsymbol{w}}(\hat{\mathrm{w}}, y)(\phi(\boldsymbol{w}, y) - \psi_Y(y))\right]. \qquad (40)\end{aligned}$$

At first glance, this approach seems impossible because ψ_Y is unknown. A second lemma is useful to develop (40).

Lemma 3: Let X be a random variable, and let $\psi_X(x)$ be its score function if f is a differentiable function satisfying

$$\lim_{|x| \to +\infty} p_X(x) f(x) = 0 \qquad (41)$$

and then

$$E[f(x)\psi_X(x)] = -E[f'(x)]. \qquad (42)$$

■

Applying this lemma to (40), we obtain

$$\frac{\partial \mathcal{E}(\boldsymbol{w})}{\partial \boldsymbol{w}} = E\left[\phi(\boldsymbol{w}, y)\frac{\partial \phi}{\partial \boldsymbol{w}}(\boldsymbol{w}, y) + \frac{\partial^2 \phi}{\partial y \partial \boldsymbol{w}}(\boldsymbol{w}, y)\right] \qquad (43)$$

where the term $\psi_Y(y)$ has disappeared. The stochastic algorithm for minimizing $\mathcal{E}(\boldsymbol{w})$ is obtained by removing the expectation

$$\begin{aligned}\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \mu_t\bigg\{&\phi[\boldsymbol{w}(t), y(t)]\frac{\partial \phi}{\partial \boldsymbol{w}}[\boldsymbol{w}(t), y(t)] \\ &+ \frac{\partial^2 \phi}{\partial y \partial \boldsymbol{w}}[\boldsymbol{w}(t), y(t)]\bigg\}. \qquad (44)\end{aligned}$$

2) Nonlinear Models: We do not give the explicit form of $\phi(\boldsymbol{w}, y)$ because many nonlinear regressors can be used. We investigated two approaches. The first is based on a linear parametric model, and we prove that our approach is equivalent to Pham's approach [29]. The second uses a multilayer perceptron that provides more flexible and general models.

In the case of a linear parametric model, ϕ can be written as

$$\phi(\boldsymbol{w}, u) = \sum_{j=1}^{k} w_j K_j(u) = \boldsymbol{w}^T \mathcal{K}(y) \tag{45}$$

denoting $\mathcal{K}(y) = (K_1(y), K_2(y), \cdots, K_k(y))^T$, where K_i, $i = 1, \cdots, k$ are nonlinear functions. Since model (45) is linear with respect to the parameters, the mean square error minimization leads to a set of k equations

$$E[\phi(\boldsymbol{w}^*, u)K_j(y) + K_j'(y)] = 0, \qquad j = 1, 2, \cdots, k \tag{46}$$

where $K_j'(y) = (dK_j/dy)(y)$. In matrix notation, (46) is expressed as

$$E[\mathcal{K}(y)\mathcal{K}^T(y)]\boldsymbol{w}^* = -E[\mathcal{K}'(y)] \tag{47}$$

whose solution, assuming $E[\mathcal{K}(y)\mathcal{K}^T(y)]$ is nonsingular, is

$$\boldsymbol{w}^* = -E[\mathcal{K}(y)\mathcal{K}^T(y)]^{-1}E[\mathcal{K}'(y)]. \tag{48}$$

These equations were first used by Pham *et al.* [29] for linear instantaneous mixtures by projecting the score function in a subspace spanned by the nonlinear functions $\mathcal{K}$. They correspond to the normal equations of the least squares problem. The linear parametric model is useful but suffers from some limitations. In particular, it requires a correct choice of the projection basis functions, i.e., some information about the source distribution.

It is also interesting to link this method with the previous approximation based on the Gram–Charlier expansion. For this purpose, we impose $K_j(u) = u^{j-1}$, $j = 1, \cdots, 4$, and (45) becomes

$$\phi(\boldsymbol{w}, u) = w_1 + w_2 u + w_3 u^2 + w_4 u^3. \tag{49}$$

Using (48), we can compute the optimal parameters $\boldsymbol{w}^*$. For even pdf with zero mean and standard deviation, the least mean square estimation provides

$$\hat{\psi}_Y^{\mathrm{MSE}}(y) = \frac{3E[y^4] - E[y^6]}{E[y^4]^2 - E[y^6]} y + \frac{E[y^4] - 3}{E[y^4]^2 - E[y^6]} y^3 \tag{50}$$

whereas the fourth-order Gram–Charlier expansion is

$$\psi_Y^{\mathrm{GC}}(y) = -\frac{E[y^4] - 1}{2} y + \frac{E[y^4] - 3}{6} y^3. \tag{51}$$

For example, if Y is uniformally distributed in the interval $[-\sqrt{3}, \sqrt{3}]$, we get

$$\hat{\psi}_Y^{\mathrm{MSE}}(y) = \tfrac{5}{2} y - \tfrac{35}{18} y^3 \tag{52}$$

$$\hat{\psi}_Y^{\mathrm{GC}}(y) = -\tfrac{2}{5} y - \tfrac{1}{5} y^3. \tag{53}$$

The plot of these functions, which is given in Fig. 5, points out large differences between the two estimators.

To overcome the classic limitations of a linear parametric model, we propose to estimate ψ_Y using a multilayer perceptron. This classic neural architecture has many interesting properties, one of which is its universal approximation capability [17]. We use a one hidden-layer perceptron with a linear output neuron

$$\phi[\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\eta}, u] = \sum_{j=1}^{k} \xi_j \sigma(w_j u - \eta_j) \tag{54}$$

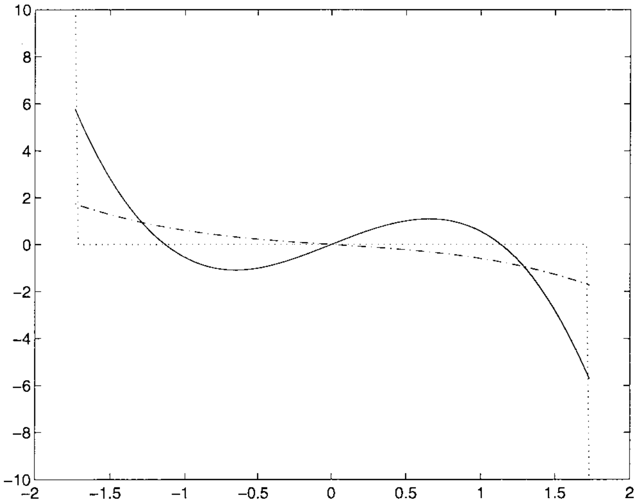


Fig. 5. Score functions of a uniform distribution. Theoretical (· · ·), ψ_Y^{GC} (--·), ψ_Y^{MSE} (-).

where σ denotes the sigmoidal activation function of the neurons. Since the model is nonlinear with respect to parameters $\boldsymbol{w}$ and $\boldsymbol{\eta}$, we cannot find a direct algebraic solution for the MSE minimization. The parameters $\boldsymbol{w}$, $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ are updated according to the adaptive algorithm (44). Denoting $z(t) = \phi[\boldsymbol{w}(t), \boldsymbol{\xi}(t), \boldsymbol{\eta}(t), y(t)]$ as the output of the neural network, (44) leads to the stochastic gradient algorithms

$$\begin{cases} \boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \mu_t \Delta \boldsymbol{w}(t) \\ \boldsymbol{\xi}(t+1) = \boldsymbol{\xi}(t) - \mu_t \Delta \boldsymbol{\xi}(t) \\ \boldsymbol{\eta}(t+1) = \boldsymbol{\eta}(t) - \mu_t \Delta \boldsymbol{\eta}(t) \end{cases} \tag{55}$$

with

$$\begin{cases} \Delta w_j(t) = \xi_j(t)\{\sigma'[w_j(t)y(t) - \eta_j(t)](1 + y(t)z(t)) \\ \qquad\qquad + w_j(t)\sigma''[w_j(t)y(t) - \eta_j(t)]\} \\ \Delta \xi_j(t) = \sigma[w_j(t)y(t) - \eta_j(t)]z(t) \\ \qquad\qquad + w_j(t)\sigma'[w_j(t)y(t) - \eta_j(t)] \\ \Delta \eta_j(t) = -\xi_j(t)\{\sigma'[w_j(t)y(t) - \eta_j(t)]z(t) \\ \qquad\qquad + w_j(t)\sigma''[w_j(t)y(t) - \eta_j(t)]\} \\ j = 1, 2, \cdots, k. \end{cases} \tag{56}$$

Note that this multilayer perceptron is trained by unsupervised learning. The gradient algorithm can be accelerated by using second-order techniques that must compute the inverse of the Hessian. A usual approximation of the Hessian is the covariance of the gradient of the regressor, the inverse of which can be iteratively calculated using the matrix inversion lemma.

3) Discussion: Recently [35], a derivation of a batch quasi-nonparametric procedure for source separation in PNL mixtures pointed out an interesting fact about the choice of score functions. In fact, it leads to two sets of estimating equations, which may be written asymptotically as

$$E[\hat{\psi}_{Y_i}(y_i)y_j] + \delta_{ij} = 0, \quad \forall i, j \tag{57}$$

$$E\left[\left\{\sum_l \hat{\psi}_{Y_l}(y_l)b_{lj}\right\}\delta(x_j - \tau_j) + \delta'(x_j - \tau_j)\right] = 0 \quad \forall j \text{ and } \forall \tau_j \in \mathbb{R}. \tag{58}$$

The first set of equations is the same as those found in the linear source separation case [29]. A good choice of the approximation of score functions must at least produce an unbiased estimator of the inverse of the PNL mixing system. A necessary condition is that the set of estimating equations must be satisfied when $y_i,\ i = 1, 2, \cdots, n$ are the true sources. For the first set of equations, this requires that $E[\hat{\psi}_i(s_i)] = 0$ if the source s_i has a nonzero mean, and nothing otherwise. For the second set, conditions on score functions are more restrictive because it must satisfy

$$\begin{aligned}\psi_{Y_j}(\tau_j) \\ = E\left[\left\{\sum_l \hat{\psi}_{S_l}\left(\sum_k b_{lk}x_k\right)b_{lj}\right\}\Big/ x_j = \tau_j\right] \quad \forall j.\end{aligned} \tag{59}$$

In other words, the approximations $\hat{\psi}_{S_l}$ of ψ_{S_l} $l = 1, 2, \cdots, n$ must be such that

$$\begin{aligned}\psi_{Y_j}(\tau_j) \\ = \arg\min_{K_j} E\left[\left\{\sum_l \hat{\psi}_{S_l}\left(\sum_k b_{lk}x_k\right)b_{lj} - K_j(x_j)\right\}^2\right]\end{aligned} \tag{60}$$

which, in general, is fulfilled neither by a GC approximation nor by a linear projection on an *ad hoc* basis of nonlinear functions.

C. Conclusion

We have shown [33] that this method can be applied in the generic problem of entropy optimization. It was also successfully applied to the linear source separation problem and leads to excellent results concerning the rejection rates. In particular, we modeled the complex source separation problem (common in narrowband communications) as an entropy optimization problem and applied the same method for complex random variables (see [3]).

VI. Algorithms and Experiments

A. Algorithms

As described in the previous sections, the separation architecture consists of two major stages: a nonlinear stage and a linear stage. In addition, at the output of the linear stage, we add score estimation blocks that estimate the output score functions that are necessary for optimal estimation of the parameters of the two stages (see Fig. 6).

We use a multilayer perceptron to model the nonlinear parametric functions $g_k(\boldsymbol{\theta}_k, e_k),\ k = 1, \cdots, n$ with linear output neuron

$$g_k(\boldsymbol{\theta}_k, u) = g[\boldsymbol{w}_k, \boldsymbol{\xi}_k, \boldsymbol{\eta}_k, u] = \sum_{j=1}^{N_k} \xi_j^k \sigma(w_j^k u - \eta_j^k) \tag{61}$$

where N_k denotes the number of hidden units of the kth MLP. Denoting the derivative of each function

$$u_k(t) = g_k'(\boldsymbol{\theta}_k(t), e_k(t)) \tag{62}$$

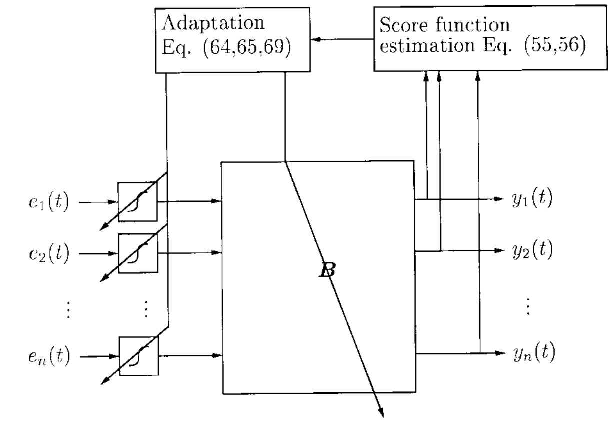


Fig. 6. Algorithm architecture.

and

$$\boldsymbol{\phi} = B^T\hat{\boldsymbol{\psi}} \tag{63}$$

the parameters of each MLP can be updated as follows:

$$j = 1, 2, \cdots, N_k$$

$$\begin{cases}\Delta\xi_j^k(t) = w_j^k \dfrac{\sigma'(w_j^k e_k - \eta_j^k)}{u_k} + \boldsymbol{\phi}_k \sigma(w_j^k e_k - \eta_j^k) \\ \Delta w_j^k(t) = \xi_j^k \dfrac{w_j^k e_k \sigma''(w_j^k e_k - \eta_j^k) + \sigma'(w_j^k e_k - \eta_j^k)}{u_k} \\ \qquad\qquad + \boldsymbol{\phi}_k \xi_j^k e_k \sigma'(w_j^k e_k - \eta_j^k) \\ \Delta\eta_j^k(t) = -\xi_j^k w_j^k \dfrac{\sigma''(w_j^k e_k - \eta_j^k)}{u_k} \\ \qquad\qquad + \boldsymbol{\phi}_k \xi_i^k \sigma'(w_j^k e_k - \eta_j^k)\end{cases} \tag{64}$$

and

$$\begin{cases}\boldsymbol{\xi}_k(t+1) = \boldsymbol{\xi}_k(t) - \mu_t \Delta\boldsymbol{\xi}_k(t) \\ w_k(t+1) = \boldsymbol{w}_k(t) - \mu_t \Delta\boldsymbol{w}_k(t). \\ \boldsymbol{\eta}_k(t+1) = \boldsymbol{\eta}_k(t) - \mu_t \Delta\boldsymbol{\eta}_k(t)\end{cases} \tag{65}$$

To deal with the indeterminacies of the nonlinear stage, we must ensure, for example, that the nonlinear stage provides a zero mean and a standard deviation (scale undeterminacy) output. This can be done with the cost function

$$E[(x_k - E[x_k])^2] + E[x_k]^2 - \log E[(x_k - E[x_k])^2] - 1 \tag{66}$$

which is always positive and vanishes if and only if

$$E[x_k] = 0 \quad \text{and} \quad E[x_k^2] = 1. \tag{67}$$

Adding this term to the expression of mutual information also leads to algorithm (64) but with

$$\boldsymbol{\phi}_k = \sum_{j=1}^n b_{jk}\hat{\psi}_j(y_j) - \left(x_i - \frac{x_i - m_i}{E[(x_i - m_i)^2]}\right) \tag{68}$$

where $m_i = E[x_i]$. In practice, the expectation will be replaced by an empirical mean or by a first-order lowpass filter that is suitable in an adaptive context.

To estimate the linear stage, i.e., matrix $\boldsymbol{B}$, we use an equivariant algorithm by a postmultiplication of (25) by $\boldsymbol{B}^T\boldsymbol{B}$,[1]

[1] The effect of $\boldsymbol{B}^T\boldsymbol{B}$ on the algorithm is treated by Cardoso *et al.* [6] and Amari *et al.* [1].

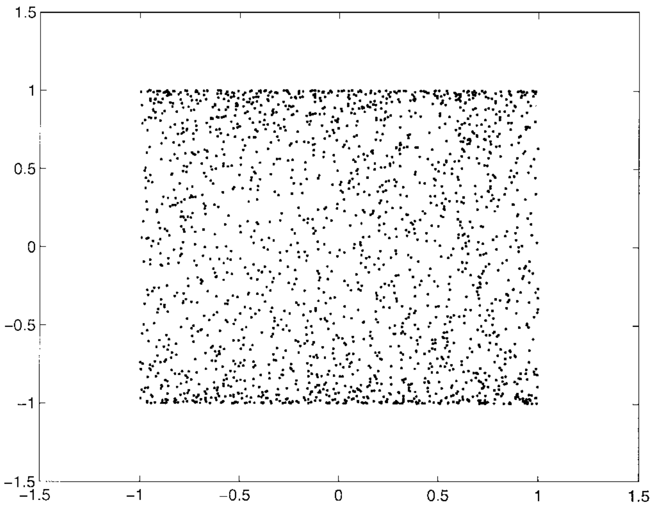


Fig. 7. Source distribution (s_2, s_1) points out the source independence.

which produces the algorithm

$$B(t+1) = (I + \mu_t \mathcal{E})B(t) \tag{69}$$

where the entries of matrix $\mathcal{E}$ are

$$\epsilon_{ij} = \begin{cases} (1-y_i^2), & \text{if } i=j \\ \hat{\psi}_{Y_i}(y_i)y_j, & \text{otherwise.} \end{cases} \tag{70}$$

The score functions used in the two algorithms are estimated by the algorithm (55) and (56).

B. Experiments

In this section, some computer simulations of the above algorithm are presented for two PNL mixtures of two sources. The independent sources are a sine wave and uniformally distributed white noise.[2] Fig. 7 shows the joint distribution of these two sources and points out the independence of the two signals. These signals are first linearly mixed with the (randomly chosen) mixture matrix

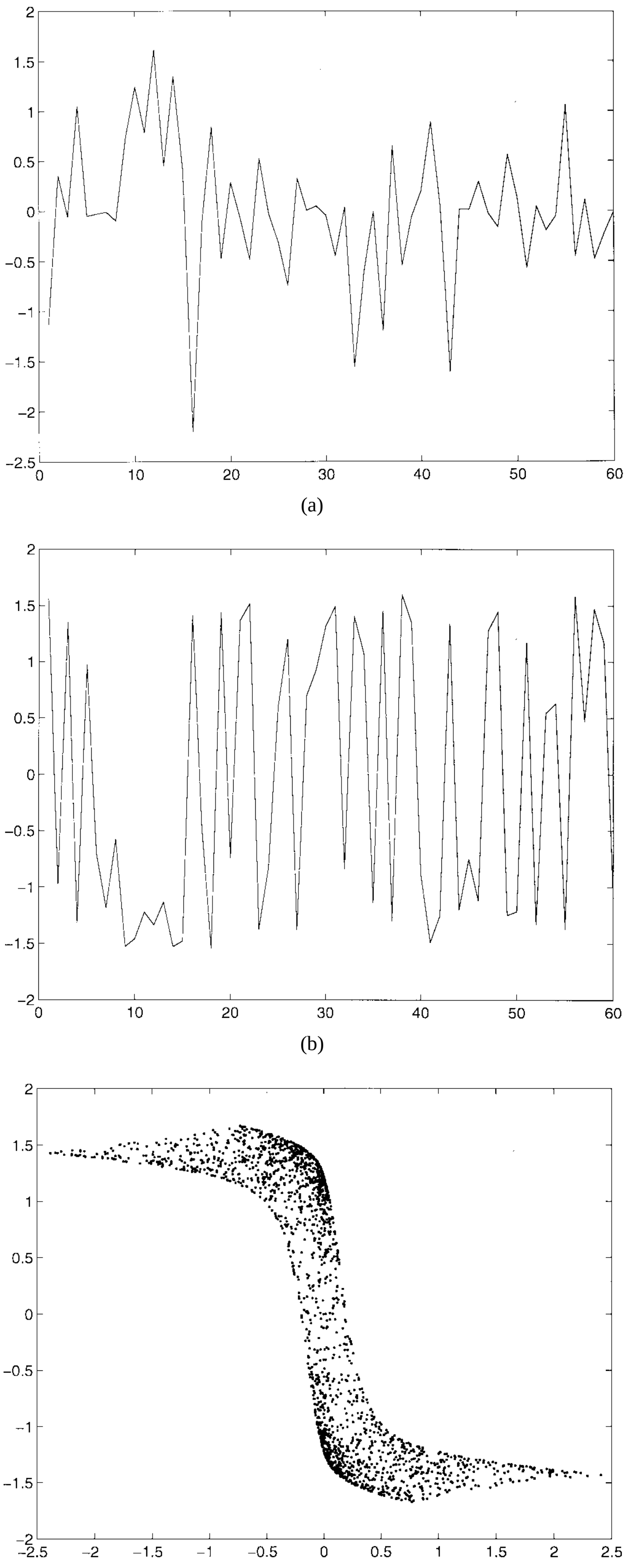
$$A = \begin{bmatrix} -2.29 & 0.49 \\ 1.84 & 0.41 \end{bmatrix}.$$

Then, nonlinear distortion

$$\begin{aligned} f_1(u) &= \tfrac{1}{10}(x+x^3), \text{ on channel 1} \\ f_2(u) &= \tfrac{3}{10}x + \tanh 3x, \text{ on channel 2} \end{aligned} \tag{71}$$

is applied to each mixture, producing a PNL mixture (see Fig. 8). The joint distribution (shown in Fig. 8) indicates the nonlinear dependence between the two signals.

The adaptive version (65) of the algorithm is applied to these mixtures. As discussed previously, the first stage of the separation mixture must cancel the nonlinear dependence.



(a)

(b)

(c)

Fig. 8. (a) Mixture signals, (b) after distortion, and (c) their distribution.

At convergence, the distribution of the nonlinear stage output is shown in Fig. 9. The distribution is contained within a simple parallelogram and is characteristic of linear mixtures of independent sources when the source pdf's are with bounded

[2] Sources are not restricted to signals with bounded pdf. Signals with exponential distributions have been used successfully.

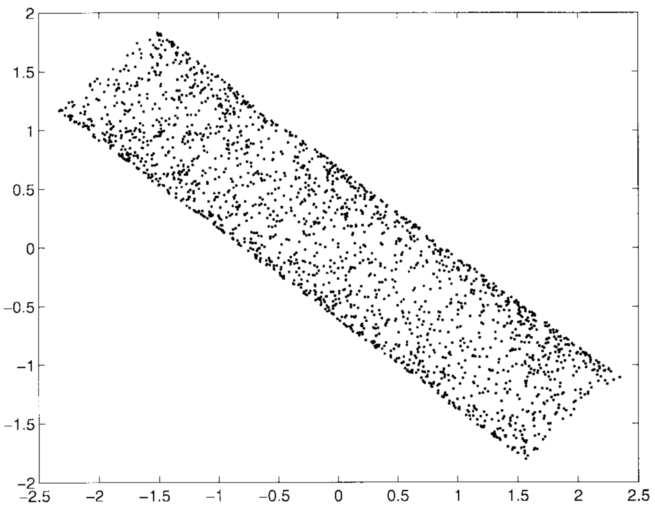


Fig. 9. Distribution after the nonlinear processing stage (x_2, x_1).

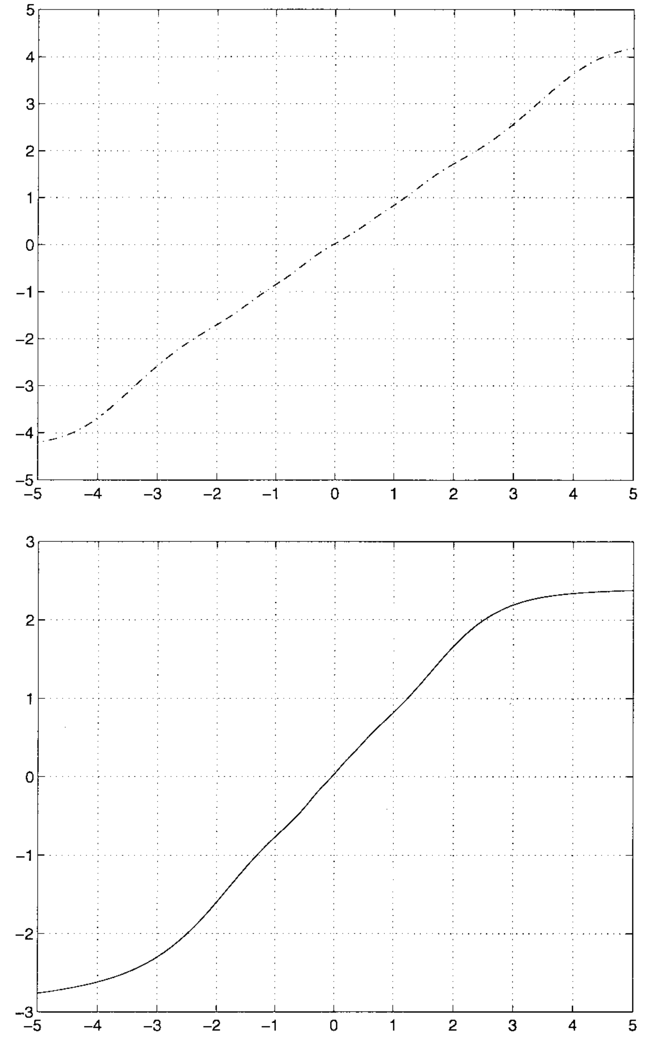
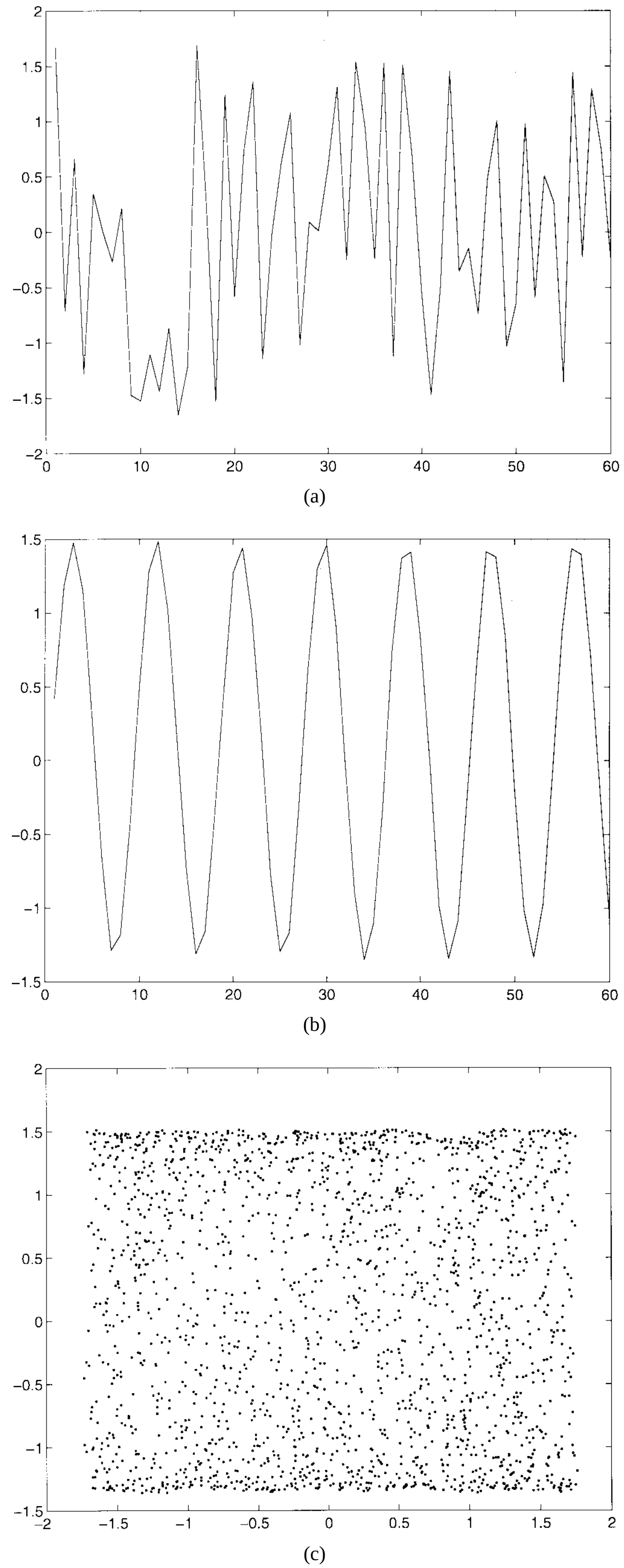


Fig. 10. Residual distortion $g_i \circ f_i$ on the two channels.



(a)

(b)

(c)

Fig. 11. (a), (b) Estimated sources $y_1(t)$ and $y_2(t)$, and (c) their joint distribution.

support. In Fig. 10, the nonlinear stage has successfully compensed the effect of the distortion and $g_i \circ f_i$ is nearly linear in the input range. The linear stage performs a classical source separation on the "*expected*" linear mixture, which is the output of the nonlinear stage. Outputs of the linear stage are shown in Fig. 11 and are good estimations of the source signals. Their joint distribution, (shown in Fig. 11) indicates that independence has been reached.

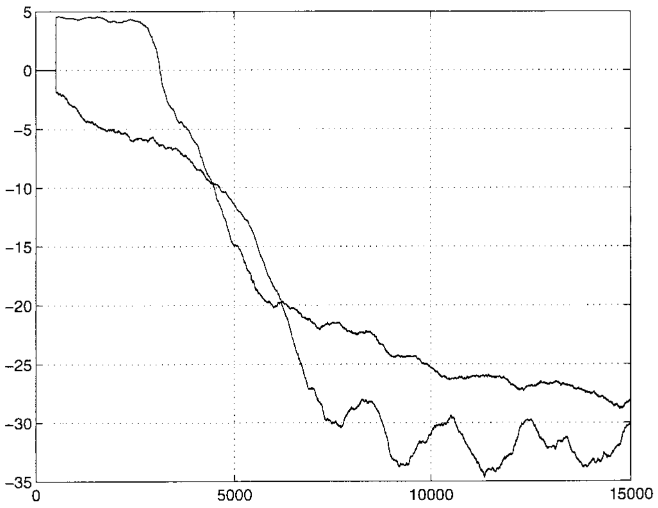


Fig. 12. Residual crosstalk (in decibels) versus iterations.

We compute the algorithm performance using the classic residual crosstalk

$$C(y, s)\text{ (dB)} = 10 \log_{10} E[(y-s)^2] \tag{72}$$

where both y and s are with unit variance. We may notice that due to the scale indeterminacy, the estimated sources and the true sources may have opposite signs. We overcome this problem by estimating the correlation coefficient between the two signals and by multiplying one of the two signals by the sign of this coefficient. The expectation operator will be replaced by an empirical expectation over a fixed number of samples. Fig. 12 shows the residual crosstalk during the algorithm run, which reaches about -30 dB at convergence. Fig. 13 shows the final estimated score functions, which are very close to the theoretical score functions.

To point out the necessity of having a nonlinear compensation stage, we illustrate in Fig. 14 the sources estimated for the same nonlinear mixture using a linear separation method. The estimated sources (Fig. 14) are very far from the true sources, and the joint distribution emphasizes their remaining statistical dependence.

VII. Conclusion

Although the general problem of source separation in nonlinear mixtures is not solvable, source separation in specific but realistic-enough mixtures, known as postnonlinear mixtures, is possible with the same indeterminacies as in linear instantaneous mixtures.

The information-based criterion used to estimate the parameters of the separating structure requires knowledge or estimation of the score functions. While poor results are achieved with a truncated estimation derived from a Gram–Charlier expansion, a direct estimation of the score functions leads to a very efficient algorithm.

Although the desired score functions are unknown, we prove that least mean square estimation (basically supervised) of the score functions is possible and leads to an unsupervised algorithm. Moreover, this method provides very good estimations of the score functions and can be used for any algorithm driven by optimization of an entropy-based criterion.

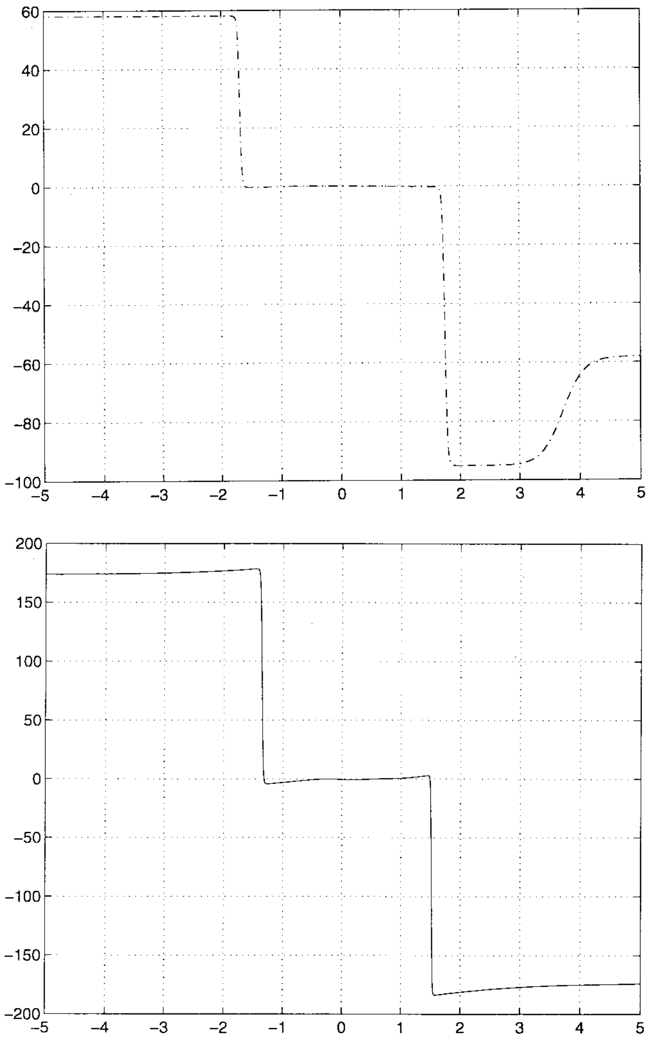


Fig. 13. Estimated score functions.

In this paper, the estimation of nonlinear stages and of (nonlinear) score functions is done using multilayer perceptrons with sigmoidal units trained by unsupervised learning. It could also be done using other nonlinear regressors, such as polynomials or splines.

Experimental results illustrate the limitations of the algorithm based on a Gram–Charlier expansion and the efficiency of the algorithm based on a direct estimation of the score functions.

We are currently studying the convergence and performance of the algorithm and the relation between the accuracy of score function estimation and separation performance with generalization to convolutive postnonlinear mixtures.

Appendix

Proof of Lemma 1: If $h_i = g_i \circ f_i$ are linear and $\boldsymbol{B}$ is a separating matrix, it is clear that the components of $\boldsymbol{y}$ are mutually independent.

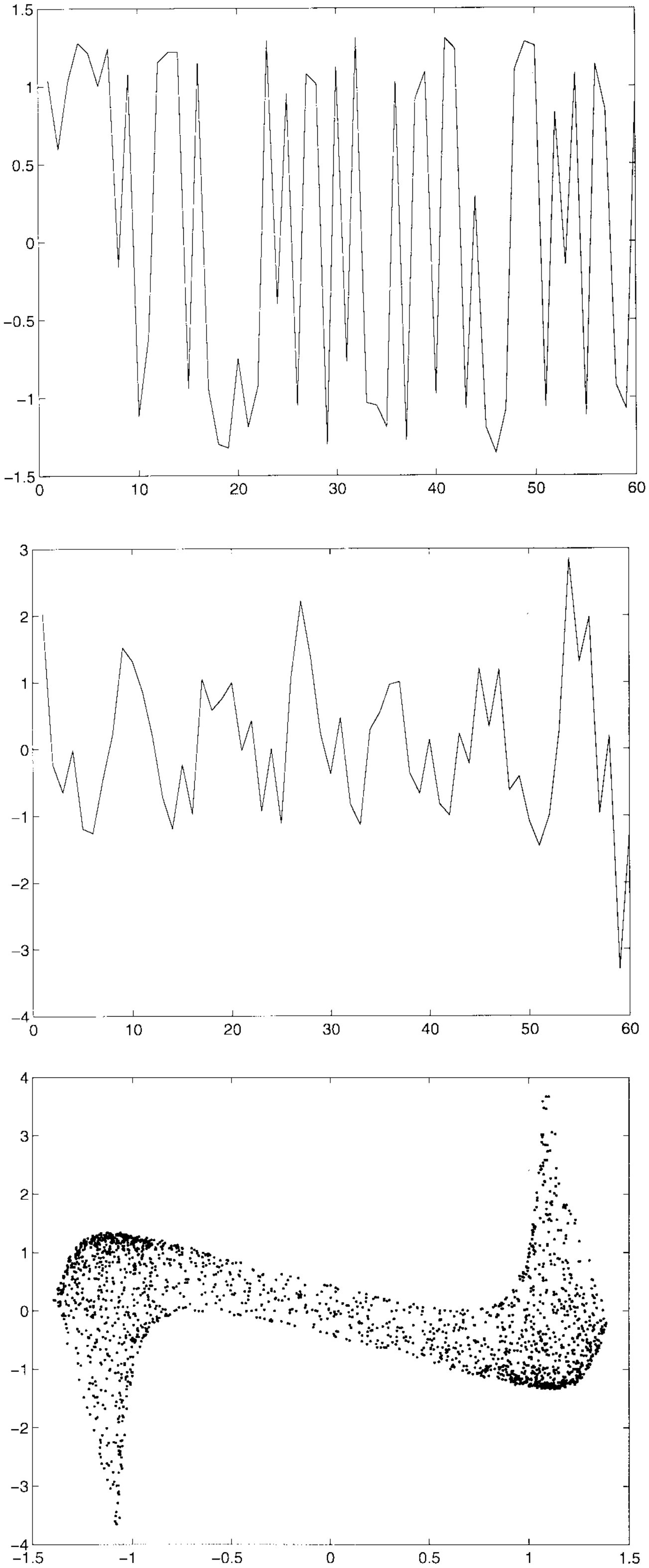


Fig. 14. Estimated sources and their joint distribution obtained using a linear separation algorithm on the PNL mixture.

Conversely, if $y_i, i = 1, 2, \cdots, n$ are mutually independent, then, denoting p_i and q_i the pdf's of s_i and y_i, respectively, the pdf of the random vector $\boldsymbol{s}$ can be written in the following two ways:

$$p_{\boldsymbol{S}}(\boldsymbol{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{73}$$

$$= \prod_{i=1}^{n} q_i\left(\sum_j b_{ij}h_j\left(\sum_k a_{jk}s_k\right)\right) \cdot \left|\prod_{i=1}^{n} h_i'\left(\sum_j a_{ij}s_j\right)\det(\boldsymbol{A})\det(\boldsymbol{B})\right| \quad \forall \boldsymbol{s} \in \mathbb{R}^n. \tag{74}$$

From assumptions, there exists a point $\boldsymbol{s}^0 \in \mathbb{R}^n$ such that $p_{\boldsymbol{S}}(\boldsymbol{s}^0) = 0$. Then, from (74) and given that $\forall i$, $\forall u \in \mathbb{R}$, $h_i'(u) \neq 0$, there exists a point $\boldsymbol{y}^0 \in \mathbb{R}^n$ such that

$$\prod_{i=1}^{n} q_i(y_i^0) = 0. \tag{75}$$

Consequently, there exists $l \in \{1, 2, \cdots, n\}$ such that

$$q_l(y_l^0) = 0. \tag{76}$$

Let, in $\mathbb{R}^n$, the hypersurface $\mathcal{H}_l$ of the implicit equation be

$$\sum_j b_{lj}h_j\left(\sum_k a_{jk}s_k\right) = y_l^0. \tag{77}$$

From (74), we have

$$\forall \boldsymbol{s} \in \mathcal{H}_l \colon p_{\boldsymbol{S}}(\boldsymbol{s}) = 0. \tag{78}$$

Points $\boldsymbol{s}$ of $\mathcal{H}_l$ are the transform of the points $\boldsymbol{x}$ of the hyperplane $\mathcal{P}_l$ of equation $\sum_j b_{lj}x_j = y_l^0$ by $\boldsymbol{A}^{-1}\boldsymbol{h}^{-1}(.)$.

Suppose that $\mathcal{H}_l$ is not parallel to any of the n hyperplanes $s_i = 0, i = 1, 2, \cdots, n$. Then, the projection of $\mathcal{H}_l$ onto each axis consists of $\mathbb{R}$. In fact, for each coordinate $s_i \in \mathbb{R}$, we can find $s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_n \in \mathbb{R}^{n-1}$ such that $\boldsymbol{h}(\boldsymbol{A}\boldsymbol{s}) \in \mathcal{P}_l$. Hence, $p_{\boldsymbol{S}}(\boldsymbol{s}) = 0$ for all $\boldsymbol{s} \in \mathbb{R}^n$, which is impossible since $p_{\boldsymbol{S}}$ sums to 1.

Consequently, for each l such that (76) holds, there exists $m_l \in \{1, 2, \cdots, n\}$ such that $\mathcal{H}_l$ is parallel to the plane $s_{m_l} = 0$. This imposes that for all $s_1, s_2, \cdots, s_n$

$$\sum_j b_{lj}h_j\left(\sum_k a_{jk}s_k\right) = k_{m_l}(s_{m_l}) \tag{79}$$

where k_{m_l} is any function. Assuming that (76) holds for all $l = 1, 2, \cdots, n$, we then have

$$\sum_j b_{lj}h_j\left(\sum_k a_{jk}s_k\right) = k_l(s_l) \tag{80}$$

where, for the sake of simplicity and without loss of generality, we set $m_l = l$. Differentiating each equation with respect to

s_i, $i = 1, 2, \cdots, n$ leads to the following matrix equation:

$$\underbrace{\begin{bmatrix} k_1'(s_1) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & k_n'(s_n)\end{bmatrix}}_{\boldsymbol{D}(\boldsymbol{s})} = \boldsymbol{B}\underbrace{\begin{bmatrix} h_1'\left(\sum_k a_{1k}s_k\right) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & h_n'\left(\sum_k a_{nk}s_k\right)\end{bmatrix}}_{\boldsymbol{W}(\boldsymbol{s})}\boldsymbol{A}. \tag{81}$$

Hence, for $\boldsymbol{s}_1 \neq \boldsymbol{s}_2$, we have

$$\boldsymbol{D}(\boldsymbol{s}_1) = \boldsymbol{B}\boldsymbol{W}(\boldsymbol{s}_1)\boldsymbol{A} \tag{82}$$

$$\boldsymbol{D}(\boldsymbol{s}_2) = \boldsymbol{B}\boldsymbol{W}(\boldsymbol{s}_2)\boldsymbol{A}. \tag{83}$$

Eliminating $\boldsymbol{B}$ leads to

$$\boldsymbol{D}(\boldsymbol{s}_1)^{-1}\boldsymbol{D}(\boldsymbol{s}_2) = \boldsymbol{A}^{-1}\boldsymbol{W}(\boldsymbol{s}_1)^{-1}\boldsymbol{W}(\boldsymbol{s}_2)\boldsymbol{A} \tag{84}$$

which we may write as

$$\boldsymbol{A}\boldsymbol{\Lambda}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \boldsymbol{\Omega}(\boldsymbol{s}_1, \boldsymbol{s}_2)\boldsymbol{A} \tag{85}$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$ are diagonal matrices. Hence, for each i, j

$$a_{ij}(\lambda_{jj}(\boldsymbol{s}_1, \boldsymbol{s}_2) - \omega_{ii}(\boldsymbol{s}_1, \boldsymbol{s}_2)) = 0. \tag{86}$$

Since $\boldsymbol{A}$ is regular, for each column j, there exists at least one $i = \sigma(j)$, where σ is a permutation, such that $a_{\sigma(j)j} \neq 0$; hence

$$\forall j\colon \lambda_{jj}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \omega_{\sigma(j)\sigma(j)}(\boldsymbol{s}_1, \boldsymbol{s}_2). \tag{87}$$

Suppose there exists another nonzero entry on the column j: $a_{\alpha(j)j} \neq 0$ and $\alpha(j) \neq \sigma(j)$; then

$$\lambda_{jj}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \omega_{\alpha(j)\alpha(j)}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \omega_{\sigma(j)\sigma(j)}(\boldsymbol{s}_1, \boldsymbol{s}_2). \tag{88}$$

Fixing $\boldsymbol{s}_2$ at a constant value and setting the $\boldsymbol{s}_1 = \boldsymbol{s}$ variable leads to

$$h'_{\alpha(j)}\left(\sum_k a_{\alpha(j),k}s_k\right) = C_j h'_{\sigma(j)}\left(\sum_k a_{\sigma(j),k}s_k\right) \tag{89}$$

where C_j is a constant. Since $\alpha(j) \neq \sigma(j)$ and $\boldsymbol{A}$ is regular, the two linear forms involved in this equation are independent, and we can write

$$h'_{\alpha(j)}(x) = C_j h'_{\sigma(j)}(y) \quad \forall x, y \in \mathbb{R}. \tag{90}$$

This shows obviously that $h_{\alpha(j)}$ and $h_{\sigma(j)}$ are linear.

Similar results are obtained by considering two nonzero entries on the rows of $\boldsymbol{A}$.

We supposed that the functions h_j and their inverses are defined over all $\mathbb{R}$. This is nonrestrictive since if an h_j was defined only on a subset $K_j \subset \mathbb{R}$, the pdf $p_{\boldsymbol{S}}$ would be necessarily null whenever $\boldsymbol{A}\boldsymbol{s} \notin \mathbb{R}^{n-1} \times K_j$. The same reasoning holds for h_j^{-1}.

Since the h_i components are linear, the problem reduces to a linear source separation problem, and $\boldsymbol{B}$ is a separating matrix.

Proof of Lemma 2: The pdf of Y can be written as

$$p_Y(y) = q(y, \boldsymbol{\gamma}) \tag{91}$$

with

$$\int q(u, \boldsymbol{\gamma})\, du = 1. \tag{92}$$

Deriving (92) with respect to $\boldsymbol{\gamma}$ leads to

$$\frac{d}{d\boldsymbol{\gamma}} \int q(u, \boldsymbol{\gamma})\, du = 0. \tag{93}$$

Using $(d/d\boldsymbol{\gamma})q(u, \boldsymbol{\gamma}) = q(u, \boldsymbol{\gamma})(d \log q(u, \boldsymbol{\gamma})/d\boldsymbol{\gamma})$, (93) becomes

$$E\left[\frac{d}{d\boldsymbol{\gamma}} \log q(y, \boldsymbol{\gamma})\right] = \int q(u, \boldsymbol{\gamma}) \frac{d}{d\boldsymbol{\gamma}} \log q(u, \boldsymbol{\gamma})\, du = 0. \tag{94}$$

Using $y = h(\boldsymbol{x}, \boldsymbol{\gamma})$, the entropy of Y can be written as

$$\begin{aligned} H(Y) &= -E[\log p_Y(y)] = -E[\log p_Y(h(\boldsymbol{\gamma}, \boldsymbol{x}))] \\ &= -E[\log q(h(\boldsymbol{\gamma}, \boldsymbol{x}), \boldsymbol{\gamma})]. \end{aligned} \tag{95}$$

Taking the derivative of (95) with respect to $\boldsymbol{\gamma}$ gives

$$\begin{aligned} \frac{d}{d\boldsymbol{\gamma}} H(Y) = &-E\left[\frac{dh(\boldsymbol{\gamma}, \boldsymbol{x})}{d\boldsymbol{\gamma}} \psi_Y(h(\boldsymbol{\gamma}, \boldsymbol{x}))\right] \\ &- E\left[\frac{d}{d\boldsymbol{\gamma}} \log q(u, \boldsymbol{\gamma})\right]. \end{aligned} \tag{96}$$

Because the second term is null according to (94), the proof is completed.

Proof of Lemma 3: The condition on f includes the existence of $E[f(x)\psi_X(x)]$ and $E[f'(x)]$. Integrating by parts the first member of (42)

$$\begin{aligned} E[f(x)\psi_X(x)] &= \int_{-\infty}^{+\infty} p_X(u)f(u)\psi_X(u)\, du \\ &= \int_{-\infty}^{+\infty} p'_X(u)f(u)\, du \\ &= [p_X(u)f(u)]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} p_X(u)f'(u)\, du \\ &= -E[f'(x)]. \end{aligned}$$

Note that the Lemma 3 can be generalized to the multivariate case (see Taleb and Jutten [33]). The score function is then the gradient of the joint pdf.

REFERENCES

[1] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, pp. 251–276, 1998.

[2] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, 1995.

[3] S. Bozinoski, A. Taleb, J.-C. Guizzo, and C. Jutten, "Séparation de sources, application la séparation de signaux et de brouilleurs dans un satellite de télécommunications," in *Proc. GRETSI*, Grenoble, France, Sept. 1997, pp. 95–98.

[4] G. Burel, "Blind separation of sources: A nonlinear neural algorithm," *Neural Networks*, vol. 5, pp. 937–947, 1992.

[5] J.-F. Cardoso, "Source separation using higher order moments," in *Proc. ICASSP*, Glasgow, U.K., May 1989, pp. 2109–2212.

[6] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 3017–3030, Dec. 1996.

[7] N. Charkani and Y. Deville, "Optimization of the asymptotic performance of time-domain convolutive source separation algorithms," in *Proc. ESANN*, Bruges, Belgium, Apr. 1997, pp. 273–278.

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

[9] G. Darmois, "Analyse des liaisons de probabilité," in *Proc. Int. Stat. Conf. 1947*, Washington, DC, 1951, vol. III A, p. 231.

[10] G. Darmois, "Analyse générale des liaisons stochastiques," *Rev. Inst. Internat. Stat.*, vol. 21, pp. 2–8, 1953.

[11] G. Deco and W. Brauer, "Nonlinear higher-order statistical decorrelation by volume-conserving architectures," *Neural Networks*, vol. 8, pp. 525–535, 1995.

[12] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Process.*, vol. 45, pp. 59–83, 1995.

[13] ———, "Séparation aveugle adptative de mélanges convolutifs," in *Proc. GRETSI*, Juan-Les-Pins, France, Sept. 1995, pp. 281–284.

[14] M. Gaeta and J.-L. Lacoume, "Source separation without *a priori* knowledge: The maximum likelihood solution," in *Proc. EUSIPCO*, Barcelona, Spain, Sept. 1990, vol. 2, pp. 621–624.

[15] A. Gorokov and P. Loubaton, "Second-order blind identification of convolutive mixtures with temporally correlated sources: A subspace based approach," in *Proc. EUSIPCO*, Trieste, Italy, Sept. 1996, pp. 2093–2096.

[16] J. Hérault, C. Jutten, and B. Ans, "Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé," in *Proc. GRETSI*, Nice, France, May 20–24, 1985, pp. 1017–1022.

[17] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no 9, pp. 359–366, 1989.

[18] C. Jutten and J. Hérault, "Blind separation of sources, Part I: An adaptive algorithm based on a neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, 1991.

[19] C. Jutten, L. N. Thi, E. Dijkstra, E. Vittoz, and J. Caelen, "Blind separation of sources: An algorithm for separation of convolutive mixtures," in *Proc. Int. Signal Process. Workshop Higher Order Stat.*, Chamrousse, France, July 1991, pp. 273–276.

[20] J. Karhunen, "Neural approaches to independent component analysis and source separation," in *Proc. ESANN*, Bruges, Belgium, Apr. 1996, pp. 249–266.

[21] M. J. Korenberg and I. W. Hunter, "The identification of nonlinear biological systems: LNL cascade models," *Biol. Cybern.*, vol. 55, pp. 125–134, 1996.

[22] J.-L. Lacoume and P. Ruiz, "Sources identification: A solution based on cumulants," in *Proc. IEEE ASSP Workshop*, Mineapolis, MN, Aug. 1988.

[23] E. Lukacs, "A characterization of the gamma distribution," *Ann. Math. Statist.*, vol. 26, pp. 319–324, 1955.

[24] A. Mansour, C. Jutten, and P. Loubaton, "Subspace method for blind separation of sources in convolutive mixtures," in *Proc. EUSIPCO*, Trieste, Italy, Sept. 1996, pp. 2081–2084.

[25] E. Moreau and O. Macchi, "New self-adaptive algorithms for source separation based on contrast functions," in *Proc. IEEE Signal Process. Workshop Higher-Order Stat.*, South Lake Tahoe, CA, June 1993, pp. 215–219.

[26] P. Pajunen, A. Hyvarinen, and J. Karhunen, "Non linear source separation by self-organizing maps," in *Proc. ICONIP*, Hong Kong, Sept. 1996, vol. 2, pp. 1207–1210.

[27] A. Parashiv-Ionescu, C. Jutten, A. M. Ionescu, A. Chovet, and A. Rusu, "High performance magnetic field smart sensor arrays with source separation," in *Proc. MSM*, Santa Clara, CA, Apr. 1998, pp. 666–671.

[28] D. T. Pham, "Séparation aveugle de sources via une analyse en composantes indépendantes," in *Proc. GRETSI*, Juan-Les-Pins, France, Sept. 1995, pp. 289–292.

[29] D. T. Pham, P. Garat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *Proc. EUSIPCO*, Brussels, Belgium, Aug. 1992, pp. 771–774.

[30] S. Prakriya and D. Hatzinakos, "Blind identification of LTI-ZMNL-LTI nonlinear channel models," *IEEE Trans. Signal Processing*, vol. 43, pp. 3007–3013, Dec. 1995.

[31] V. P. Skitovic, "Linear forms of independent random variables and the normal distribution law," *Izvestiya Akademii Nauk SSSR. Seriya Matematiceskaya*, vol. 18, pp. 185–200, 1954, in Russian.

[32] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, 5th ed. Oxford, U.K.: Oxford Univ. Press, 1991, vol. 1.

[33] A. Taleb and C. Jutten, "Entropy optimization, application to blind source separation," in *Proc. ICANN*, Lausanne, Switzerland, Oct. 1997, pp. 529–534.

[34] ———, "Nonlinear source separation: The post-nonlinear mixtures," in *Proc. ESANN*, Bruges, Belgium, Apr. 1997, pp. 279–284.

[35] ———, "Batch algorithm for source separation in postnonlinear mixtures," in *Proc. ICA*, Aussois, France, Jan. 1999, pp. 155–160.

[36] H. L. N. Thi and C. Jutten, "Blind sources separation for convolutive mixtures," *Signal Process.*, vol. 45, pp. 209–229, 1995.

[37] S. Van Gerven and D. Van Compernolle, "Feedforward and feedback in a symmetric adaptive noise canceller: Stability analysis in a simplified case," in *Proc. EUSIPCO*, Brussels, Belgium, Aug. 1992, pp. 1081–1084.

[38] L. Xu, C. C. Cheung, and S. Amari, "Nonlinearity and separation capability: Further justification for the ica algorithm with mixtures of densities," in *Proc. ESANN*, Bruges, Belgium, Apr. 1997, pp. 291–296.

[39] H. H. Yang, S. Amari, and A. Cichocki, "Information back-propagation for blind separation of sources from nonlinear mixture," in *Proc. ICNN*, Houston, TX, 1996.

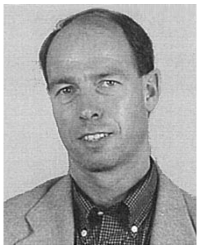
[40] ———, "Information-theoritic approach to blind separation of sources in nonlinear mixture," *Signal Process.*, vol. 64, no. 3, pp. 291–300, 1998.

[41] H. H. Yang and S. I. Amari, "Adaptive on-line learning algorithms for blind separation—Maximum entropy and minimum mutual information," *Neural Comput.*, vol. 9, no. 7, pp. 1457–1482, 1997.

[42] D. Yellin and E. Weinstein, "Multichannel signal separation: Methods and analysis," *IEEE Trans. Signal Processing*, vol. 44, pp. 106–118, Jan. 1996.

Anisse Taleb received the electrical engineering diploma from the ENSIEG-INPG, Grenoble, France, in 1996. He is currently a Ph.D. student within the LIS-INPG Laboratory, Grenoble.

His research interests include statistical signal processing, neural networks, and nonlinear source separation.



Christian Jutten received the Ph.D. degree in 1981 and the D.Sc. degree in 1987 from the Institut National Polytechnique, Grenoble, France.

He taught as an Associate Professor at the Ecole Nationale Suprieure d'Electronique et de Radiolectricit, Grenoble, from 1982 to 1989. He was Visiting Professor with the Ecole Polytrechnique Federale de Lausanne, Lausanne, Switzerland, in 1989 before he became a Full Professor with the Universit Joseph Fourier, Grenoble. For 15 years, his research interests have been learning in neural networks, source separation, and independent component analysis.