Markovian Source Separation

Shahram Hosseini, Christian Jutten, *Associate Member, IEEE*, and Dinh Tuan Pham, *Member, IEEE*

***Abstract*—A maximum likelihood (ML) approach is used to separate the instantaneous mixtures of temporally correlated, independent sources with neither preliminary transformation nor *a priori* assumption about the probability distribution of the sources. A Markov model is used to represent the joint probability density of successive samples of each source. The joint probability density functions are estimated from the observations using a kernel method. For the special case of autoregressive models, the theoretical performance of the algorithm is computed and compared with the performance of second-order algorithms and i.i.d.-based separation algorithms.**

***Index Terms*—Independent component analysis, Markov process, maximum likelihood, source separation, temporal correlation.**

I. Introduction

In this work, the maximum likelihood (ML) approach is used for blind separation of linear instantaneous mixtures of independent sources. In a general framework (without noise and with same number of sensors and sources), this problem can be formulated as follows. Having N samples of K instantaneous mixtures of K sources, $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$, where $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_K(t)]^T$ and $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_K(t)]^T$ are, respectively, the vectors of the observations and of the sources, and $\mathbf{A}$ is the mixing matrix, one wants to find an estimation of the matrix $\mathbf{A}$ (or of its inverse, the separation matrix) up to a scaling and a permutation. Many methods have been proposed [1]–[11]; most of them pay no attention to the time structure of data.

It is known that the time structure of data may be used for improving the estimation of the model [12]–[15]. This additional information can actually make the estimation of the model possible in cases where the basic independent component analysis (ICA) methods can not estimate it, for example, if the sources are Gaussian but correlated over time. Moreover, most of the methods exploiting the time structure are second-order methods that are basically simpler than higher order statistics methods. These methods try to diagonalize the time-lagged covariance matrix using one [12], [13] or several [14]–[17] time lags. A good review can be found in [18, ch. 18]. The second-order approaches provide generally the unbiased estimators, but as we will see, the estimation is not efficient unless for the special case of the Gaussian sources.

One of the separation approaches consists of maximizing the likelihood function of the observations. This approach has the advantage of providing an estimator asymptotically efficient (minimum variance among unbiased estimators). For the i.i.d. sources, this method has been used by Pham and Garat [9]. They show that the separation matrix can be estimated by solving the system of equations $E[y_i\psi_j(y_j)] = 0, \forall i \neq j$, where y_i represents the estimation of the ith source, and $\psi_j(.)$ is the score function of the jth source. In the same paper, the authors propose another method, for temporally correlated sources, that consists of computing the discrete Fourier transform (DFT) of the sources and in applying the ML approach on the results. In [19], the authors use also the ML method, but they model the probability densities of the sources using a fourth-order truncated Gram–Charlier development. In [20], the ML method is used to separate the Gaussian sources where the correlation of each source is modeled by an autoregressive model. Finally, [21] studies a general theory of estimating functions of independent component analysis when the independent source signals are temporally correlated and considers the ML method for estimating the separating matrix.

In this work, we study the problem in the case of temporally correlated sources, and our objective is to maximize directly the likelihood function without either any preliminary transformation or *a priori* assumption concerning the probability density of the sources. In fact, these densities will be estimated during the maximization procedure with a kernel approach.

The paper is organized as follows. In Section II, after the problem statement, we derive the likelihood function to be maximized, and we show its equivalence with a conditional mutual information minimizing algorithm. In Section III, we propose an iterative equivariant algorithm for estimating the separation matrix and discuss practical issues, especially a method for estimating the conditional score functions. In Section IV, the theoretical performance of the algorithm for the special case of autoregressive (AR) source models is computed, and some interesting conclusions are derived. The simulation results with both artificial and real-world data are presented in Section V. Finally, in Section VI, we conclude and present the perspectives.

II. Theory

Having N samples of a K-dimensional vector $\mathbf{x}$ resulting from a linear transformation $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{s}$ is a vector of independent signals, eventually correlated in the time (the sources), and $\mathbf{A}$ is a $K \times K$ invertible matrix, our objective is to estimate the separating matrix $\mathbf{B} = \mathbf{A}^{-1}$ up to classical indeterminacies, using an ML approach.

Manuscript received May 31, 2002; revised March 18 2003. This work was supported in part by the European project BLInd Source Separation and applications (BLISS, IST 1999-14190). The associate editor coordinating the review of this paper and approving it for publication was Dr. Hamid Krim.

S. Hosseini was with LIS-INPG, Grenoble, France. He is now with the Laboratoire d'Acoustique, Metrologie, Instrumentation, 31062 Toulouse, France (e-mail: hosseini@cict.fr).

C. Jutten is with the Laboratoire des Images et des Signaux, UMR CNRS 5083, INPG, UJF, Grenoble, France (e-mail: Christian.Jutten@inpg.fr).

D. T. Pham is with the Laboratoire de Modélisation et Calcul, Grenoble, France (e-mail: Dinh-Tuan.Pham@imag.fr).

Digital Object Identifier 10.1109/TSP.2003.819000

A. ML Method

The ML method consists of maximizing the joint probability density function (pdf) of all the samples of all the components of the vector $\mathbf{x}$ (all the observations), with respect to $\mathbf{B}$. We denote this pdf as

$$f(x_1(1),\ldots,x_K(1),\ldots\ldots,x_1(N),\ldots,x_K(N)). \quad (1)$$

Under the assumption of independence of the sources, this function is equal to

$$\left(\frac{1}{|\det(\mathbf{B}^{-1})|}\right)^N \prod_{i=1}^{K} f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(1),\mathbf{e}_i^T\mathbf{B}\mathbf{x}(2),\ldots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(N)) \quad (2)$$

where $f_{s_i}(.)$ represents the joint density of N samples of the source s_i, and $\mathbf{e}_i$ is the ith column of the identity matrix. We suppose now that the sources are qth-order Markov sequences, i.e.,

$$f_{s_i}(s_i(t)|s_i(t-1),\ldots,s_i(1)) = f_{s_i}(s_i(t)|s_i(t-1),\ldots,s_i(t-q)). \quad (3)$$

Using (3), (2) reduces to (4), shown at the bottom of the page. Taking the logarithm of (4), one obtains the log-likelihood function that must be maximized to estimate the separating matrix $\mathbf{B}$, as in (5), shown at the bottom of the page. Using the Bayes formula, we can replace the conditional densities by the joint densities. After the simplification, the function to be maximized becomes (6), shown at the bottom of the page. Until now, we supposed that the source density functions were known, but this is not actually the case. Thus, the true maximum likelihood approach may consist of parametrizing these densities and in maximizing the parametrized likelihood function with respect to parameters. This approach is not, however, applicable because in absence of *a priori* knowledge on the sources, it is not possible to parametrize their densities correctly. Therefore, the densities must be estimated using a nonparametric approach. Since the sources are not observable, their densities could be estimated only via the reconstructed sources y_i. Thus, the functions f_{s_i} in (6) must be replaced with the estimations of the density functions of reconstructed sources f_{y_i}.

B. Minimization of Conditional Mutual Information

We can also study the problem from another point of view: minimization of the conditional mutual information of the estimated sources $\mathbf{y} = \mathbf{B}\mathbf{x}$ with respect to the separating matrix $\mathbf{B}$. For the qth-order Markov sequences, the qth-order conditional mutual information can be defined by (7), shown at the bottom of the page, which is always non-negative, and zero if and only if the processes $y_i(t)$ are statistically independent for $i = 1,\ldots,K$[22]. Using the expectation operator $E[.]$, we can write

$$I = E[\log f_{\mathbf{y}}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q))] - \sum_{i=1}^{K} E[\log f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q))] \quad (8)$$

which can be rewritten as

$$I = E[\log f_{\mathbf{x}}(\mathbf{x}(t)|\mathbf{x}(t-1),\ldots,\mathbf{x}(t-q))] - \log|\det(\mathbf{B})| - \sum_{i=1}^{K} E[\log f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q))]. \quad (9)$$

$$\left(\frac{1}{|\det(\mathbf{B}^{-1})|}\right)^N \prod_{i=1}^{K}\left[f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(1),\ldots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(q)) \prod_{t=q+1}^{N} f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t)|\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1),\ldots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-q))\right]. \quad (4)$$

$$L_1 = N\log(|\det(\mathbf{B})|) + \sum_{i=1}^{K}\left[\log(f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(1),\ldots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(q))) + \sum_{t=q+1}^{N}\log(f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t)|\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1),\ldots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-q)))\right]. \quad (5)$$

$$L_1 = N\log(|\det(\mathbf{B})|) + \sum_{i=1}^{K}\left[\sum_{t=q+1}^{N}\log(f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t),\ldots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-q))) - \sum_{t=q+2}^{N}\log(f_{s_i}(\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-1),\ldots,\mathbf{e}_i^T\mathbf{B}\mathbf{x}(t-q)))\right]. \quad (6)$$

$$I = \int_{\mathcal{R}^K} f_{\mathbf{y}}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q))\log\frac{f_{\mathbf{y}}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q))}{\prod_{i=1}^{K} f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q))}d\mathbf{y} \quad (7)$$

Because the first term is independent of $\mathbf{B}$, the separation matrix can be estimated by minimizing

$$L_2 = -\log|\det(\mathbf{B})| - \sum_{i=1}^{K} E[\log f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q))]. \quad (10)$$

In practice, under the ergodicity conditions, the mathematical expected value could be replaced by a time averaging. Having N time samples, the above criterion is rewritten as

$$L_2 = -\log|\det(\mathbf{B})| - \sum_{i=1}^{K} \frac{1}{N-q} \sum_{t=q+1}^{N} \log f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q)). \quad (11)$$

Comparing (11) with (5), it can be seen that L_2 is asymptotically equivalent to $-L_1/N$ if the actual conditional pdf of the sources f_{s_i} are replaced by the conditional pdf of the estimated sources f_{y_i}. As we mentioned in the previous subsection, this is the only practical way to use the ML method. Then, the equivalence of the ML method with the mutual information minimization method, which has already been shown for the i.i.d. signals in [23] and [24], also holds for the Markovian sources.

C. Estimating Equations

To estimate the matrix $\mathbf{B}$, we need to compute the gradient of the criterion (10) with respect to $\mathbf{B}$

$$\frac{\partial L_2}{\partial \mathbf{B}} = -\mathbf{B}^{-T} - E\left[\frac{\partial}{\partial \mathbf{B}} \sum_{i=1}^{K} \log f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q))\right]. \quad (12)$$

Since $y_j(t) = \sum_{k=1}^{K} b_{jk} x_k(t)$ depends only on the jth row of $\mathbf{B}$, i.e., on b_{jk}, $k = 1,\ldots,K$, we have

$$\begin{aligned}
&\frac{\partial}{\partial b_{jk}} \sum_{i=1}^{K} \log f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q)) \\
&= \frac{\partial}{\partial b_{jk}} \log f_{y_j}(y_j(t)|y_j(t-1),\ldots,y_j(t-q)) \\
&= \sum_{l=0}^{q} \frac{\partial \log f_{y_j}(y_j(t)|y_j(t-1),\ldots,y_j(t-q))}{\partial y_j(t-l)} \frac{\partial y_j(t-l)}{\partial b_{jk}} \\
&= \sum_{l=0}^{q} \frac{\partial \log f_{y_j}(y_j(t)|y_j(t-1),\ldots,y_j(t-q))}{\partial y_j(t-l)} x_k(t-l). \quad (13)
\end{aligned}$$

Definition 1: Suppose $z_0,\ldots,z_q$ are $q+1$ random variables with joint pdf $f_z(z_0,\ldots,z_q)$. The conditional score function of z_0 given $z_1,\ldots,z_q$, which is denoted by $\psi_z(z_0|z_1,\ldots,z_q)$, is defined as the gradient of the function $-\log f_z(z_0|z_1,\ldots,z_q)$, where $f_z(z_0|z_1,\ldots,z_q)$ is the conditional pdf of z_0 given $z_1,\ldots,z_q$.

Note that the conditional score function is a vector of size $q+1$. Its lth component is denoted by $\psi_z^{(l)}(z_0|z_1,\ldots,z_q) = -(\partial/\partial z_l)\log f_z(z_0|z_1,\ldots,z_q)$. Using this definition, (13) can be rewritten as

$$\begin{aligned}
&\frac{\partial}{\partial b_{jk}} \sum_{i=1}^{K} \log f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q)) \\
&= -\sum_{l=0}^{q} \psi_{y_j}^{(l)}(y_j(t)|y_j(t-1),\ldots,y_j(t-q)) x_k(t-l). \quad (14)
\end{aligned}$$

Denote $\boldsymbol{\psi}_{\mathbf{y}}^{(l)}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q))$ as the column vector of size K with general component $\psi_{y_j}^{(l)}(y_j(t)|y_j(t-1),\ldots,y_j(t-q))$ and $\mathbf{x}(t-l) = [x_1(t-l),\ldots,x_K(t-l)]^T$

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{B}} \sum_{i=1}^{K} \log f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q)) \\
&= -\sum_{l=0}^{q} \boldsymbol{\psi}_{\mathbf{y}}^{(l)}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q)) \cdot \mathbf{x}^T(t-l). \quad (15)
\end{aligned}$$

Thus

$$\frac{\partial L_2}{\partial \mathbf{B}} = -\mathbf{B}^{-T} + E\left[\sum_{l=0}^{q} \boldsymbol{\psi}_{\mathbf{y}}^{(l)}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q)) \cdot \mathbf{x}^T(t-l)\right]. \quad (16)$$

Note that if $q = 0$, we retrieve the classical result for i.i.d. sources [9]. Solving $(\partial L_2/\partial \mathbf{B}) = 0$ with respect to $\mathbf{B}$ yields

$$E\left[\sum_{l=0}^{q} \boldsymbol{\psi}_{\mathbf{y}}^{(l)}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q)).\mathbf{x}^T(t-l)\right] = \mathbf{B}^{-T}. \quad (17)$$

Post-multiplying the above equation by $\mathbf{B}^T$, we obtain

$$E\left[\sum_{l=0}^{q} \boldsymbol{\psi}_{\mathbf{y}}^{(l)}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q)).\mathbf{y}^T(t-l)\right] = \mathbf{I}. \quad (18)$$

This yields the $K(K-1)$ estimating equations

$$E\left[\sum_{l=0}^{q} \psi_{y_i}^{(l)}(y_i(t)|y_i(t-1),\ldots,y_i(t-q)) y_j(t-l)\right] = 0 \quad i \neq j = 1,\ldots,K \quad (19)$$

which determine $\mathbf{B}$ up to a scaling and a permutation. The other K equations

$$E\left[\sum_{l=0}^{q} \psi_{y_i}^{(l)}(y_i(t)|y_i(t-1),\ldots,y_i(t-q)) y_i(t-l)\right] = 1 \quad i = 1,\ldots,K$$

are not important and can be replaced by any other scaling convention.

The system of equations (19) may be solved using, for example, the Newton–Raphson adaptive algorithm. However, in the paper, we preferred to minimize directly the criterion (10) using (16) in a gradient descent scheme because its realization is more straightforward. The drawback of the gradient method is that its performance depends on the choice of the learning rate parameter. A bad choice of this parameter may lead to divergence.

III. Algorithm

In this section, we propose an iterative equivariant algorithm to estimate the separating matrix using the method proposed in the previous section. Since the realization of the algorithm requires the estimation of the conditional score functions of the estimated sources, we study this problem first.

A. Estimation of the Conditional Score Functions

For estimating the conditional score functions, we can first estimate the conditional densities and derive then the conditional score functions by computing the gradient of their logarithms. The estimation of the conditional densities may be done using the estimation of the joint pdf of $q+1$ successive samples of each source by a kernel method, which is very time consuming. It must be also noticed that the distribution of the data in the $(q+1)$th dimensional space is not symmetric because of the temporal correlation between the samples. Thus, one should either use the nonsymmetrical kernels or apply a prewhitening transformation on data. In the first versions of our work [25], we only considered $q=1$, and we used the Fukunaga formula [26] to estimate the two-dimensional (2-D) joint densities. At first, this approach prewhitens the data by linearly transforming them to have a unit covariance matrix; next, it smoothes the data by using a symmetrical kernel, and finally, it transforms back the data. The method was highly time consuming even for 2-D data in the case of first-order Markovian sources.

Recently, Pham [27] proposed another algorithm to compute the conditional score functions. The method starts with a prewhitening stage to obtain noncorrelated temporal data. Pham also suggests that the time prewhitening can allow a reduction of the dimension of the used kernels because a great part of the independence between the variables is evacuated. The influence of the prewhitening on the estimation of the score functions is computed and will later be compensated using an additive term. Afterwards, the joint entropies of whitened data are estimated using a discrete Riemann sum and the third-order cardinal spline kernels. The conditional entropies, which are defined as

$$H(y_i(t)|y_i(t-1),\ldots,y_i(t-q)) = -E[\log f_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q))] \quad (20)$$

are computed by estimating the joint entropies

$$H(y_i(t)|y_i(t-1),\ldots,y_i(t-q)) = H(y_i(t),y_i(t-1),\ldots,y_i(t-q)) - H(y_i(t-1),\ldots,y_i(t-q)). \quad (21)$$

The estimator $\hat{H}(y_i(t)|y_i(t-1),\ldots,y_i(t-q))$ is a function of the observations $y_i(1),\ldots,y_i(N)$. The lth component of the conditional score function in a sample point $y_i(n)$ is computed as

$$\hat{\psi}^{(l)}_{y_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q))|_{t=n} = N\frac{\partial \hat{H}(y_i(t)|y_i(t-1),\ldots,y_i(t-q))}{\partial y_i(n-l+1)}. \quad (22)$$

The method is very powerful and provides quite a good estimation of the conditional score functions.

B. Equivariant Iterative Algorithm

The estimation of the separating matrix $\mathbf{B}$ is done using a batch iterative approach. At each iteration, using the current value of the matrix $\mathbf{B}$, the conditional score functions of the estimated sources are estimated, and the gradient (16) is computed. Afterwards, the matrix $\mathbf{B}$ is updated to minimize the criterion (10) using a relative gradient descent scheme to achieve an equivariant estimation [28]:

$$\mathbf{B}_{\text{new}} = \left(\mathbf{I} - \mu\frac{\partial L_2(\mathbf{B})}{\partial \mathbf{B}}\mathbf{B}^T_{\text{old}}\right)\mathbf{B}_{\text{old}}. \quad (23)$$

Using (16)

$$\mathbf{H} = \frac{\partial L_2(\mathbf{B})}{\partial \mathbf{B}}\mathbf{B}^T_{\text{old}} = -\mathbf{I} + E\left[\sum_{l=0}^{q}\boldsymbol{\psi}^{(l)}_{\mathbf{y}}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q))\cdot\mathbf{y}^T(t-l)\right]. \quad (24)$$

Because of the scaling indeterminacy, the diagonal entries of the matrix $\mathbf{H}$ have no importance. Thus, we can replace $\mathbf{H}$ by only the second term on the right-hand side of (24), which is denoted

$$\mathbf{G} = E\left[\sum_{l=0}^{q}\boldsymbol{\psi}^{(l)}_{\mathbf{y}}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q))\cdot\mathbf{y}^T(t-l)\right]. \quad (25)$$

Hence, the update formula becomes

$$\mathbf{B}_{\text{new}} = (\mathbf{I} - \mu\mathbf{G})\mathbf{B}_{\text{old}}. \quad (26)$$

To remove the ambiguity due to the scaling indeterminacy, the rows of the separating matrix $\mathbf{B}$ are normalized at each iteration so that the estimated sources have unit variance.

IV. Statistical Properties of the ML Estimator for the Special Case of Autoregressive Models

In this section, we compute the bias and the variance of the ML estimator for a special case of qth-order Markovian sources, i.e., when the sources are generated by the qth-order autoregressive models:

$$s_i(t) = \sum_{k=1}^{q}\rho_{ki}s_i(t-k) + n_i(t) \qquad i=1,\ldots,K \quad (27)$$

where $n_i(t)$ are the i.i.d. sequences. In this case, the conditional densities (3) become

$$f_{s_i}(s_i(t)|s_i(t-1),\ldots,s_i(t-q)) = f_{n_i}\left(s_i(t)-\sum_{k=1}^{q}\rho_{ki}s_i(t-k)\right) = f_{n_i}\left(-\sum_{k=0}^{q}\rho_{ki}s_i(t-k)\right) \quad (28)$$

where $\rho_{0i}=-1$. At convergence, the estimated sources y_i are in the proximity of the actual sources. Thus, we can suppose that each y_i also matches the autoregressive model (27) so that

$$f_{s_i}(y_i(t)|y_i(t-1),\ldots,y_i(t-q)) = f_{n_i}\left(-\sum_{k=0}^{q}\rho_{ki}y_i(t-k)\right). \quad (29)$$

Thus, the components of the conditional score functions are

$$\psi_{y_i}^{(l)}(y_i(t)|y_i(t-1),\ldots,y_i(t-q)) = -\rho_{li}\psi_{n_i}\left(-\sum_{k=0}^{q}\rho_{ki}y_i(t-k)\right) \quad i=1,\ldots,K \;\; 0\le l\le q \tag{30}$$

and the estimating equations (19) become (by replacing the mathematical expected value $E[.]$ with the time average $E_N[.]$)

$$E_N\left[\psi_{n_i}\left(-\sum_{k=0}^{q}\rho_{ki}y_i(t-k)\right)\left(-\sum_{l=0}^{q}\rho_{li}y_j(t-l)\right)\right]=0 \qquad i\neq j=1,\ldots,K. \tag{31}$$

In the derivation of the estimating equations, we used the ML approach. Thus, the above score functions are, in principle, the score functions of the generating i.i.d. signals n_i. However, it may be noted that at this stage, it is possible to relax this starting hypothesis and replace the score functions by any arbitrary function (denoted by $\psi_i(.)$ in the following) in the estimating equations (31). In the following subsection, we are interested in the separating equations for the functions ψ_i, which are chosen arbitrarily. We show that the estimator obtained by solving these equations is unbiased for a large class of functions, and we compute the asymptotic variance of the estimator in the general case. It should, however, be noted that the optimal ML estimator, asymptotically providing the minimum variance, can be achieved only by using the actual score functions. We will study the properties of this optimal estimator at the end of this section.

A. Case of Arbitrarily Chosen Estimating Functions

We consider now the estimating equations (31), in which the score functions $\psi_{n_i}(.)$ are replaced by the arbitrarily chosen scalar functions $\psi_i(.)$. To compute the bias and the asymptotic covariance matrix of the estimation error, we adapt the method used by Pham and Garat [9] for i.i.d. sources. If the matrices $\mathbf{A}$ and $\mathbf{B}$ are normalized with the same convention, one may expect that $\mathcal{E}=\mathbf{I}-\mathbf{BA}$ is small and compute the first-order Taylor expansion of the estimating equations (31) around $\mathbf{A}$. The relation $\mathbf{B}=\mathbf{A}^{-1}-\mathcal{E}\mathbf{A}^{-1}$ implies

$$y_i(t)=s_i(t)-\sum_{k=1}^{K}\varepsilon_{ik}s_k(t) \tag{32}$$

and therefore, we have (33), shown at the bottom of the page, which can be rewritten, considering (27), as (34), shown at the bottom of the page, in which the signal $v_j(t)=-\sum_{l=0}^{q}\rho_{li}s_j(t-l)$ can be interpreted as the source j filtered by the whitening filter of the source i.[1] The first-order Taylor expansion of (34) gives (35), shown at the bottom of the page, which yields, neglecting the second-order terms, (36), shown at the bottom of the page. As $N\to\infty$, the temporal means $E_N[\psi_i(n_i(t))s_k(t-l)]$ and $E_N[\psi'(n_i(t))v_j(t)s_k(t-l)]$ converge to the mathematical expectations $E[\psi_i(n_i(t))s_k(t-l)]$ and $E[\psi'(n_i(t))v_j(t)s_k(t-l)]$, which vanishes unless $k=i$ or $k=j$ respectively. Thus, (36) becomes

$$\begin{aligned}&E_N[\psi_i(n_i(t))v_j(t)]\\ &\approx\varepsilon_{ji}E\left[\psi_i(n_i(t))\left(-\sum_{l=0}^{q}\rho_{li}s_i(t-l)\right)\right]\\ &\quad+\varepsilon_{ij}E\left[\psi_i'(n_i(t))v_j(t)\left(-\sum_{l=0}^{q}\rho_{li}s_j(t-l)\right)\right]\\ &\quad i\neq j=1,\ldots,K\end{aligned} \tag{37}$$

[1]In the sense that the filter applied to $s_i(t)$ provides $n_i(t)$.

$$E_N\left[\psi_i\left(-\sum_{l=0}^{q}\rho_{li}\left[s_i(t-l)-\sum_{k=1}^{K}\varepsilon_{ik}s_k(t-l)\right]\right)\left(-\sum_{l=0}^{q}\rho_{li}\left[s_j(t-l)-\sum_{k=1}^{K}\varepsilon_{jk}s_k(t-l)\right]\right)\right]=0 \qquad i\neq j=1,\ldots,K \tag{33}$$

$$E_N\left[\psi_i\left(n_i(t)+\sum_{l=0}^{q}\rho_{li}\sum_{k=1}^{K}\varepsilon_{ik}s_k(t-l)\right)\left(v_j(t)+\sum_{l=0}^{q}\rho_{li}\sum_{k=1}^{K}\varepsilon_{jk}s_k(t-l)\right)\right]=0 \qquad i\neq j=1,\ldots,K \tag{34}$$

$$E_N\left[\left(\psi_i(n_i(t))+\psi_i'(n_i(t))\sum_{l=0}^{q}\rho_{li}\sum_{k=1}^{K}\varepsilon_{ik}s_k(t-l)\right)\left(v_j(t)+\sum_{l=0}^{q}\rho_{li}\sum_{k=1}^{K}\varepsilon_{jk}s_k(t-l)\right)\right]=0 \qquad i\neq j=1,\ldots,K \tag{35}$$

$$E_N[\psi_i(n_i(t))v_j(t)]=-\sum_{l=0}^{q}\rho_{li}\sum_{k=1}^{K}\{\varepsilon_{jk}E_N[\psi_i(n_i(t))s_k(t-l)]+\varepsilon_{ik}E_N[\psi_i'(n_i(t))v_j(t)s_k(t-l)]\} \quad i\neq j=1,\ldots,K. \tag{36}$$

which can be rewritten as

$$E_N[\psi_i(n_i(t))v_j(t)] \approx \varepsilon_{ji}E[\psi_i(n_i(t))n_i(t)] + \varepsilon_{ij}E[\psi_i'(n_i(t))v_j^2(t)] \quad i \neq j = 1, \ldots, K. \quad (38)$$

Our objective is to compute the mean and the covariance matrix of the off-diagonal entries of the error matrix $\mathcal{E}$.

Definition 2: The off-diagonal terms of the error matrix $\mathcal{E}$, i.e., $\varepsilon_{ij} \quad i \neq j = 1, \ldots, K$, are structured in $K(K-1)/2$ elementary vectors

$$\boldsymbol{\delta}_{(ij)} := \begin{bmatrix} \varepsilon_{ij} \\ \varepsilon_{ji} \end{bmatrix} \qquad i \neq j = 1, \ldots, K \quad (39)$$

which are concatenated in a vector $\boldsymbol{\delta}$ of size $K(K-1)$

$$\boldsymbol{\delta} := [\cdots \boldsymbol{\delta}_{(ij)}^T \cdots]^T.$$

In the same way, one defines

$$\mathbf{g}_{(ij)} := \begin{bmatrix} E_N[\psi_i(n_i(t))v_j(t)] \\ E_N[\psi_j(n_j(t))v_i(t)] \end{bmatrix} \qquad i \neq j = 1, \ldots, K$$

and

$$\mathbf{g} := [\cdots \mathbf{g}_{(ij)}^T \cdots]^T.$$

We now want to compute the mean and the variance of the error vector $\boldsymbol{\delta}$. We can rewrite (38) in the following form:

$$\mathbf{g} = \mathbf{H}\boldsymbol{\delta} \quad (40)$$

in which $\mathbf{H}$ is a bloc diagonal matrix with blocs

$$\mathbf{H}_{(ij)} = \begin{pmatrix} E[\psi_i'(n_i(t))]E[v_j^2(t)] & E[n_i(t)\psi_i(n_i(t))] \\ E[n_j(t)\psi_j(n_j(t))] & E[\psi_j'(n_j(t))]E[v_i^2(t)] \end{pmatrix}. \quad (41)$$

The covariance matrix of the vector $\mathbf{g}$ is $\mathbf{G} = E[\mathbf{g}\mathbf{g}^T]$. To compute this matrix, we must compute the terms of the form $E[E_N[\psi_i(n_i(t))v_j(t)]E_N[\psi_k(n_k(t))v_l(t)]]$ for $i \neq j$ and $k \neq l$. If $\psi_i(n_i(t))$ are zero mean,[2] the only nonzero terms of $\mathbf{G}$ are those with $\{i,j\} = \{k,l\} \quad i \neq j = 1, \ldots, K$. Thus, the covariance matrix $\mathbf{G}$ is bloc diagonal with blocs as in (42), shown at the bottom of the page, which can be rewritten as (43), shown at the bottom of the page. $n_i(t)$ and $n_i(r)$ are independent if $t \neq r$ because $n_i(t)$ is an i.i.d. signal. If $E[\psi_i(n_i(t))] = 0$, as mentioned above, then the two diagonal entries of the summand of (43) are zero unless $t = r$. What about the off-diagonal entries? Considering the definition of $v_i(t)$, this signal only depends on the past and present values of the signal $n_i(t)$. Therefore, if $t > r$, then $E[\psi_i(n_i(t))v_i(r)] = 0$, and if $t < r$, then $E[\psi_i(n_i(r))v_i(t)] = 0$. Thus, the off-diagonal entries of the summand are also zero unless $t = r$. It follows that we have (44), shown at the bottom of the page, or, after simplification, we have (45), shown at the bottom of the page. If the matrix $\mathbf{H}$ is invertible,[3] one can write, from (40) $\boldsymbol{\delta} = \mathbf{H}^{-1}\mathbf{g}$. The covariance matrix of the error vector $\boldsymbol{\delta}$ is

$$\boldsymbol{\Lambda} = E[\boldsymbol{\delta}\boldsymbol{\delta}^T] = \mathbf{H}^{-1}E[\mathbf{g}\mathbf{g}^T]\mathbf{H}^{-T} = \mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-T}. \quad (46)$$

The estimator is unbiased because

$$E[\boldsymbol{\delta}] = \mathbf{H}^{-1}E[\mathbf{g}] = \mathbf{0}. \quad (47)$$

[2]This is the case, for example, if $n_i(t)$ have symmetric distributions and ψ_i are odd functions.

[3]It is not true if $s_i(t)$ and $s_j(t)$ are two Gaussian sources, with the same spectral densities, i.e., $\rho_{li} = \rho_{lj} \quad \forall l$. In this case, $\mathbf{H}_{(ij)} = \begin{pmatrix} \sigma_{n_j}^2 & \sigma_{n_i}^2 \\ \sigma_{n_j}^2 & \sigma_{n_i}^2 \end{pmatrix}$.

$$\mathbf{G}_{(ij)} = E\begin{pmatrix} E_N[\psi_i(n_i(t))v_j(t)]E_N[\psi_i(n_i(t))v_j(t)] & E_N[\psi_i(n_i(t))v_j(t)]E_N[\psi_j(n_j(t))v_i(t)] \\ E_N[\psi_j(n_j(t))v_i(t)]E_N[\psi_i(n_i(t))v_j(t)] & E_N[\psi_j(n_j(t))v_i(t)]E_N[\psi_j(n_j(t))v_i(t)] \end{pmatrix} \quad (42)$$

$$\mathbf{G}_{(ij)} = \frac{1}{(N-q)^2}\sum_{t=q+1}^{N}\sum_{r=q+1}^{N}\begin{pmatrix} E[\psi_i(n_i(t))\psi_i(n_i(r))]E[v_j(t)v_j(r)] & E[\psi_i(n_i(t))v_i(r)]E[\psi_j(n_j(r))v_j(t)] \\ E[\psi_i(n_i(r))v_i(t)]E[\psi_j(n_j(t))v_j(r)] & E[\psi_j(n_j(t))\psi_j(n_j(r))]E[v_i(t)v_i(r)] \end{pmatrix} \quad (43)$$

$$\mathbf{G}_{(ij)} = \frac{1}{(N-q)^2}\sum_{t=q+1}^{N}\begin{pmatrix} E[\psi_i^2(n_i(t))]E[v_j^2(t)] & E[\psi_i(n_i(t))v_i(t)]E[\psi_j(n_j(t))v_j(t)] \\ E[\psi_i(n_i(t))v_i(t)]E[\psi_j(n_j(t))v_j(t)] & E[\psi_j^2(n_j(t))]E[v_i^2(t)] \end{pmatrix} \quad (44)$$

$$\mathbf{G}_{(ij)} = \frac{1}{N-q}\begin{pmatrix} E[\psi_i^2(n_i(t))]E[v_j^2(t)] & E[\psi_i(n_i(t))v_i(t)]E[\psi_j(n_j(t))v_j(t)] \\ E[\psi_i(n_i(t))v_i(t)]E[\psi_j(n_j(t))v_j(t)] & E[\psi_j^2(n_j(t))]E[v_i^2(t)] \end{pmatrix}. \quad (45)$$

B. Optimal Separation

Although the estimator is unbiased for a large class of functions ψ_i, it is not optimal in the maximum likelihood sense, unless the separating functions ψ_i are the score functions, which are denoted ψ_{n_i}. In that case, the estimator is asymptotically efficient (the variance is minimum). In this subsection, we proceed to compute the covariance matrix $\mathbf{\Lambda}$ for this optimal case. We will use the following property of the score function:

$$E[g(u)\psi(u)] = E[g'(u)] \tag{48}$$

for any differentiable function g such that $g(u)f(u)$ vanished at infinity ($f(u)$ represents the pdf of the variable u). It follows that

$$E[\psi(u)] = 0 \tag{49}$$
$$E[\psi(u)u] = 1 \tag{50}$$
$$E[\psi'(u)] = E[\psi^2(u)]. \tag{51}$$

Hence, the diagonal blocs of the matrix $\mathbf{H}$ can be written in the following form:

$$\mathbf{H}_{(ij)} = \begin{pmatrix} \sigma_{v_j}^2 E[\psi_{n_i}^2(n_i(t))] & 1 \\ 1 & \sigma_{v_i}^2 E[\psi_{n_j}^2(n_j(t))] \end{pmatrix}. \tag{52}$$

To simplify the matrix $\mathbf{G}$, we know that

$$\begin{aligned} E[\psi_{n_i}(n_i(t))v_i(t)] &= E\left[-\psi_{n_i}(n_i(t))\sum_{l=0}^{q}\rho_{lj}s_i(t-l)\right] \\ &= E[\psi_{n_i}(n_i(t))s_i(t)] \\ &= E\left[\psi_{n_i}(n_i(t))(n_i(t)+\sum_{l=1}^{q}\rho_{li}s_i(t-l))\right] \\ &= E[\psi_{n_i}(n_i(t))n_i(t)] = 1 \end{aligned} \tag{53}$$

because $n_i(t)$ is independent of $s_i(t-l)$ for $l > 0$. Thus

$$\mathbf{G}_{(ij)} = \frac{1}{N-q}\begin{pmatrix} E[\psi_i^2(n_i(t))]\sigma_{v_j}^2 & 1 \\ 1 & E[\psi_j^2(n_j(t))]\sigma_{v_i}^2 \end{pmatrix}. \tag{54}$$

From (46) and after some computations, the covariance matrix of the ML estimator error $\mathbf{\Lambda}$ is obtained, which is a bloc diagonal matrix with the diagonal blocs

$$\begin{aligned} \mathbf{\Lambda}_{(ij)} = \frac{1}{N-q}&\frac{1}{E[\psi_{n_i}^2(n_i(t))]E[\psi_{n_j}^2(n_j(t))]\sigma_{v_i}^2\sigma_{v_j}^2 - 1} \\ &\cdot\begin{pmatrix} \sigma_{v_i}^2 E[\psi_{n_j}^2(n_j(t))] & -1 \\ -1 & \sigma_{v_j}^2 E[\psi_{n_i}^2(n_i(t))] \end{pmatrix}. \end{aligned} \tag{55}$$

C. Some Theoretical Results for First-Order AR Sources

We want now to evaluate the error covariance matrix $\mathbf{\Lambda}$ in the special case of first-order autoregressive sources ($q = 1$). We suppose that all the i.i.d. signals $n_i(t)$ have the same distribution taken from the family of the generalized Gaussian densities, which are defined by

$$p(n_i(t)) = \frac{\beta}{2\sigma}\frac{\Gamma^{1/2}\left(\frac{3}{\beta}\right)}{\Gamma^{3/2}\left(\frac{1}{\beta}\right)}\exp\left\{-\left(\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}\right)^{\beta/2}\left|\frac{n_i(t)}{\sigma}\right|^{\beta}\right\} \tag{56}$$

where σ is the standard deviation, and β is a shape parameter. For $\beta = 2$, one retrieves the Gaussian law, for $\beta = 1$, the bilateral exponential law, and for $\beta \to \infty$, the uniform law. It can be easily verified that

$$E[\psi_{n_i}^2(n_i(t))] = \frac{\beta_i^2}{\sigma_{n_i}^2}\frac{\Gamma\left(\frac{3}{\beta_i}\right)}{\Gamma^2\left(\frac{1}{\beta_i}\right)}\Gamma\left(\frac{2\beta_i - 1}{\beta_i}\right). \tag{57}$$

In the following, we suppose that $\beta_i = \beta, \forall i$, and the correlated sources $s_i(t)$ have unit variance so that $\sigma_{n_i}^2 = 1 - \rho_i^2$. It can also be easily shown that

$$\begin{aligned} \sigma_{v_1}^2 &= \frac{1 + \rho_2^2 - 2\rho_2\rho_1}{1 - \rho_1^2}\sigma_{n_1}^2 \\ \sigma_{v_2}^2 &= \frac{1 + \rho_1^2 - 2\rho_2\rho_1}{1 - \rho_2^2}\sigma_{n_2}^2. \end{aligned} \tag{58}$$

Considering the relations (55), (57), and (58), the diagonal blocs of the covariance matrix $\mathbf{\Lambda}$ are

$$\mathbf{\Lambda}_{(ij)} = \frac{1}{(N-1)(1-C^2A_iA_j)}\begin{pmatrix} -CA_i\frac{1-\rho_i^2}{1-\rho_j^2} & 1 \\ 1 & -CA_j\frac{1-\rho_j^2}{1-\rho_i^2} \end{pmatrix} \tag{59}$$

where

$$\begin{aligned} C &= \beta^2\Gamma\left(\frac{2\beta-1}{\beta}\right)\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma^2\left(\frac{1}{\beta}\right)} \\ A_i &= \frac{1 + \rho_j^2 - 2\rho_i\rho_j}{1 - \rho_i^2} \\ A_j &= \frac{1 + \rho_i^2 - 2\rho_i\rho_j}{1 - \rho_j^2}. \end{aligned} \tag{60}$$

We call *total variance* the sum of the diagonal entries of $\mathbf{\Lambda}$, which writes using (39) and (46):

$$V = \sum_{i=1}^{K(K-1)}\mathbf{\Lambda}_{(ii)} = \sum_{i,j,i\neq j}E[\varepsilon_{ij}^2]. \tag{61}$$

It can be remarked that the total variance is asymptotically inversely proportional to the number of samples,and that for the particular case $\rho_i = \rho, \forall i$, it does not depend on ρ. This means that the optimal separation performance for K i.i.d. sources (of same densities) is equal to the performance for K correlated sources with the same spectral densities.

It should be also remarked that considering (32) and neglecting $\varepsilon_{ii}s_i$ in comparison with s_i, we can write

$$y_i(t) \simeq s_i(t) - \sum_{j\neq i}\varepsilon_{ij}s_j(t). \tag{62}$$

If the sources have unit variances, it is clear that

$$V \simeq \sum_{i=1}^{K}E[(y_i(t) - s_i(t))^2]. \tag{63}$$

Fig. 1 shows the total variance for the case of two first-order autoregressive sources of same generalized Gaussian law as a function of β, for $\rho_1 = 0$ and ρ_2 taking three values 0, 0.5, and 0.99. Considering the figure, the following remarks may be noted.

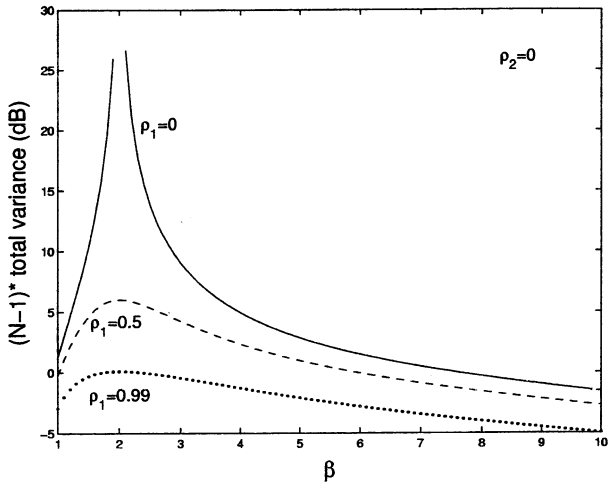


Fig. 1. Theoretical total variance of the estimator as a function of β for three values of ρ_1: 0 (solid line), 0.5 (dashed line), and 0.99 (dotted line), where ρ_2 is fixed to zero. $\beta = 2$ corresponds to Gaussian distribution.

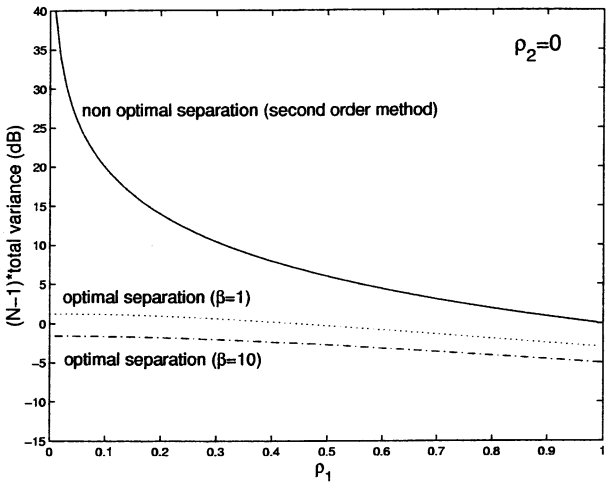


Fig. 2. Comparison of the theoretical total variance of a second-order estimator with an optimal estimator for $\beta = 1$ and $\beta = 10$ as a function of ρ_1, where ρ_2 is fixed to 0. The second-order method is only optimal for $\beta = 2$, i.e., for Gaussian sources.

1) The optimal separation performance increases when the spectral densities of the two sources become more different. This effect exists whatever the nature of the sources, but it is more remarkable for the nearly Gaussian sources.
2) The performance decreases when the sources approach the Gaussianity. The separation of two i.i.d. Gaussian sources, or two correlated Gaussian sources, provided by two i.i.d. Gaussian sources filtered by the same filter is theoretically impossible because the variance of the estimator approaches infinity.

For the particular case of two first-order autoregressive Gaussian sources of unit variance, with the correlation coefficients ρ_1 and ρ_2, the total variance is equal to

$$V = \left(\frac{1}{N-1}\right) \times \frac{(1+\rho_2^2-2\rho_1\rho_2)(1-\rho_1^2)+(1+\rho_1^2-2\rho_1\rho_2)(1-\rho_2^2)}{(1+\rho_2^2-2\rho_1\rho_2)(1+\rho_1^2-2\rho_1\rho_2)-(1-\rho_1^2)(1-\rho_2^2)}. \quad (64)$$

We now suppose that the score functions of the sources are not known, and one does not want to estimate them. In this case, it is possible to use the arbitrarily chosen functions ψ_i in the estimating equations (31). The separation is no longer optimal because the variance of the nonoptimal estimator is greater than that obtained by using the score functions that are associated with the ML estimator. A particular interesting case is when the score functions of the Gaussian sources are used to separate the non-Gaussian temporally correlated sources. This means that one only considers the second-order statistics of the sources, which brings us to the second-order algorithms, which have been largely studied in ICA literature [12]–[14], [15]–[17].

Thus, we must normally replace $\psi_i(x)$ by x (the score function of a unit variance Gaussian source) in (41) and (45) and compute the covariance matrix using the general relation (46). It is, however, easy to see that in this case, the covariance matrix will be equal to the optimal covariance matrix for the Gaussian sources so that the total variance can be computed using (64). The comparison of the total variance for $\beta = 2$ with the other values of β in Fig. 1 can give a vision of the performance gain of an optimal estimator with respect to a second-order method.

Fig. 2 compares the optimal total variance, computed by (59), with the total variance of a second-order method, computed by (64), as a function of ρ_1, where ρ_2 is fixed to zero, for different values of β. It should be remarked that a second-order method is not able to separate the sources having the same spectral densities, even if they are non-Gaussians.

V. Experimental Results

In this section, some experimental results with artificial and real-world signals will be presented.

A. Does the Algorithm Work Well?

The first experiment consists of comparing the theoretical end experimental separation performances for the special case of two AR1 Gaussian sources with respective coefficients ρ_1 and ρ_2. Two experiments are done. In the first one, we suppose that all the statistical properties of the sources are known. Thus, we suppose an AR1 Gaussian model with known model coefficients for the sources and solve the estimating equations (31) with these hypotheses. In the second experiment, we suppose that nothing is known about the sources, except that they are the first-order Markov sequences, and the equivariant algorithm of Section III-B is used for separating. Each of the experiments is done using 100 Monte Carlo simulations with $N = 1000$. At each simulation, the nondiagonal entries of the matrix $\mathcal{E} = \mathbf{I} - \hat{\mathbf{B}}\mathbf{A}$ are computed, where $\mathbf{A} = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$ is the mixing matrix, and $\hat{\mathbf{B}}$ is the estimation of the separating matrix. Then, the total variance of the estimator,V is computed using (61).

Fig. 3 shows $10\log_{10}(V * (N-1))$, for $\rho_1 = 0.5$, as a function of ρ_2. The curve is the theoretical variance, which is computed from (64). The asterisks and the circles represent the

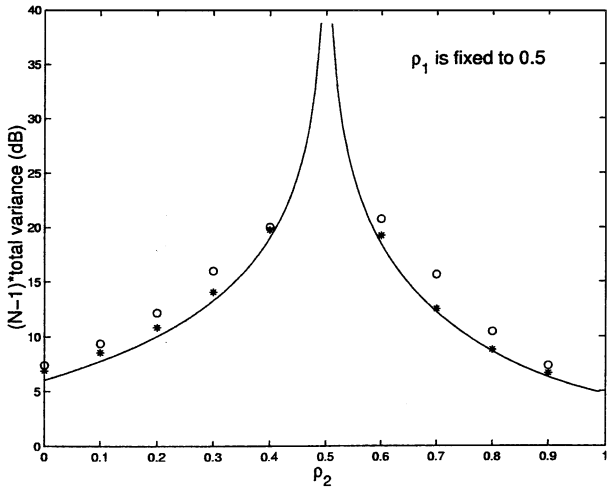


Fig. 3. Total variance of the estimator as a function of ρ_2, where ρ_1 is fixed to 0.5, for two Gaussian sources. Curve: Theoretical variance. Asterisks: Empirical variance supposing that the statistical properties of the sources are known. Circles: Empirical variance without *a priori* hypotheses about the sources.

values obtained using the first and second experiments, respectively. As can be seen, the practical results are in agreement with the theory.

B. Is Any Improvement Obtained by Taking Into Account the Temporal Correlation?

In another experiment, we compare the performance of our algorithm with the ML algorithm proposed by Pham and Garat [9], which pays no attention to the temporal correlation of data. Our objective is so to know how much improvement may be obtained by taking into account the temporal correlation. Two sources are generated by filtering two i.i.d. uniform signals using two similar AR1 filters ($\rho_1 = \rho_2$). Each of the experiments is done using 100 Monte Carlo simulations with $N = 1000$.

Fig. 4 shows $10\log_{10}(V * (N-1))$ as a function of $\rho_1 = \rho_2$. The solid and dashed lines represent, respectively, our algorithm and the Pham-Garat algorithm. The performance of our algorithm is approximately insensitive to the coefficient of AR1 filters. This is not surprising because, as we saw in Section IV-C, it is not the temporal correlation itself that provides the additional information but the difference between the spectral densities of the sources. The separation performance of our method for two i.i.d. sources is equal to the performance for two highly correlated sources with the same spectral densities. On the other hand, the performance of the Pham-Garat algorithm decreases with $\rho_1 = \rho_2$ because the filtered sources approach Gaussianity so that the separation becomes more difficult. Our algorithm works better for $\rho_1 = \rho_2 > 0.1$. For nearly uncorrelated sources, the error involved in the estimation of the conditional score functions results in a lower performance of our algorithm with respect to the Pham-Garat algorithm. Finally, note that a second-order method could not separate the sources because they have the same spectral densities.

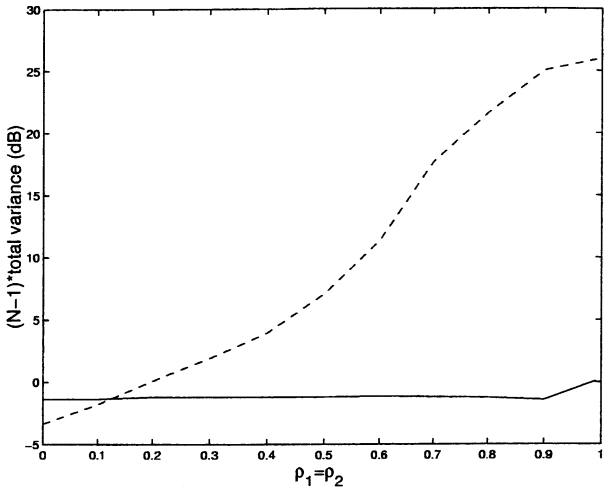


Fig. 4. Total variance of the two estimators as a function of $\rho_1 = \rho_2$ for two correlated sources obtained by AR1 filtering of two uniform i.i.d. signals. Solid line: Our algorithm. Dashed line: ML algorithm of Pham and Garat, which pays no attention to the temporal correlation of data.

C. Does the Algorithm Work Better Than a Second-Order Method?

In another experiment, we want to compare the performance of our algorithm with a second order method i.e., the AMUSE algorithm [12]. This simple and fast algorithm works as follows.

1) Whiten the (zero-mean) data $\mathbf{x}(t)$ to obtain $\mathbf{z}(t)$.
2) Compute the eigenvalue decomposition of $\overline{\mathbf{C}_1^{\mathbf{z}}} = (1/2)[\mathbf{C}_1 + \mathbf{C}_1^T]$, where $\mathbf{C}_1 = E[\mathbf{z}(t)\mathbf{z}(t-1)]$ is the first-lagged covariance matrix.
3) The rows of the separating matrix $\mathbf{B}$ are given by the eigenvectors.

The algorithm works well when the first lagged covariances are different for all the sources. For a first-order autoregressive model, which we use in our experiences, this condition is satisfied if the model coefficients ρ_i are different.

Two sources are generated by passing two i.i.d. uniform signals into two AR1 filters. Each of the experiments is done using 100 Monte Carlo simulations with $N = 1000$. Fig. 5 shows $10\log_{10}(V * (N-1))$, for $\rho_2 = 0$, as a function of ρ_1. The results confirm the theoretical curves of Fig. 2. The experiment is also performed using the Pham-Garat algorithm, and the result is shown in the same figure, which confirms the discussion of the previous subsection.

D. Experiences Using Real-World Data

In the last experiment, we use the artificial instantaneous linear mixtures of the speech signals. Our algorithm (supposing a first- or a second-order Markov model for signals), the Pham–Garat algorithm, and the AMUSE algorithm are used to separate the mixtures.

Two 4-s independent speech signals are mixed by the mixing matrix $\mathbf{A} = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$. The mixtures are divided into 220 frames of 200 samples. Thus, each frame (of length 18,2

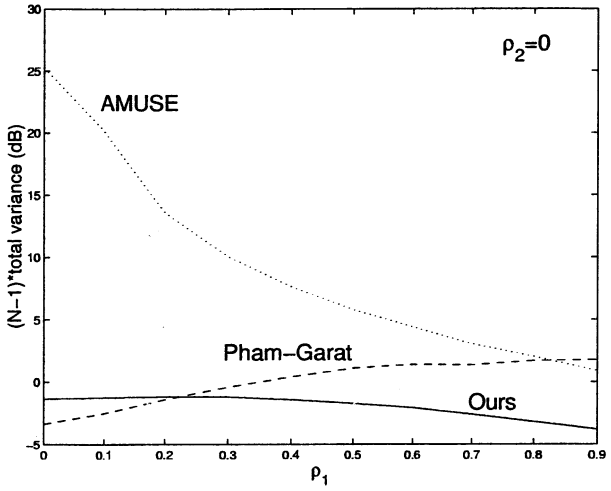


Fig. 5. Total variance of three estimators as a function of ρ_1, ρ_2 being fixed to 0. Two correlated sources are obtained by AR1 filtering of two uniform i.i.d. signals. Solid line: Our algorithm. Dotted line: AMUSE algorithm. Dashed line: ML algorithm of Pham and Garat, which pays no attention to the temporal correlation of data.

TABLE I
NORMALIZED TOTAL VARIANCE AND THE RESIDUAL CROSS-TALKS ON THE TWO CHANNELS (ALL IN DECIBELS) FOR OUR ALGORITHM (USING FIRST-ORDER AND SECOND-ORDER MARKOV MODELING OF THE SOURCES), THE PHAM-GARAT ALGORITHM, AND THE AMUSE ALGORITHM IN THE EXPERIMENT WITH SPEECH SIGNALS

Method	$10\log_{10}(V*(N-1))$	C1	C2
2nd order Markov	5.5171	-37.9169	-31.0492
1st order Markov	6.8624	-36.7051	-30.4426
Pham-Garat	9.4721	-28.0380	-25.7875
AMUSE	11.0436	-29.2890	-25.3705

ms) can be considered as a stationary sequence. Note, however that, contrary to the previous simulations, the first-order Markov model is no longer an exact model for temporally correlation of the sources.

For each method, $10\log_{10}(V*(N-1))$ is computed. We also compute the residual cross-talks (in decibels) on the two channels, which are defined as

$$C_i = 10\log_{10} E[(y_i - s_i)^2] \quad i = 1, 2 \tag{65}$$

where y_i and s_i have unit variance. The results are shown in Table I. As can be seen, our algorithm works significantly better than the two others. The second-order Markov modeling results in a better performance with respect to the first-order model, although the improvement is not considerable.

VI. CONCLUSION

In this paper, we used the maximum likelihood approach for the blind separation of the instantaneous mixtures of the temporally correlated sources with no preliminary transformation nor *a priori* assumption on the probability densities of the sources. We suppose only that the sources are the Markov sequences. The kernel estimators are used to estimate the conditional densities of the sources from the observations using an iterative algorithm. For the special case of autoregressive source models, the theoretical performance of the algorithm is computed. The experimental results confirm the relevance of the approach.

Several points could, however, be improved. The algorithm is rather slow because estimating the probability densities is time consuming. We are currently working on less expensive, faster, and more efficient algorithms to estimate these densities. The simple gradient algorithm used in this paper is sensitive to the learning rate μ. A conjugate gradient algorithm, for example, could solve this problem.

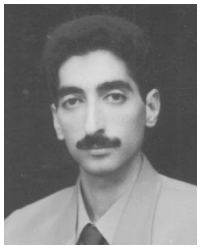
It must be noted that the computational complexity of the algorithm is inherently much higher than the complexity of suboptimal algorithms that do not pay attention to time structure of data or make some *a priori* assumptions about the probability densities of the sources. Thus, its application seems to be limited to offline ICA problems where the separation performance could be more important than computational complexity.

Other tests seem necessary to compare our algorithm with other existing algorithms. One can test the algorithm on the correlation generated by nonlinear filters. We mention that in the problem formulation, no hypothesis about the nature of temporal filters is made, except for a Markov model that simplifies the realization. The idea may be also used to separate the nonlinear and post nonlinear mixtures of temporally correlated sources.

REFERENCES

[1] 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, Mai 1985, pp. 1017–1022.
[2] J.-L. Lacoume and P. Ruiz, "Sources identification: A solution based on cumulants," in *Proc. IEEE ASSP Workshop*, Mineapolis, MN, Aug. 1988.
[3] J.-F. Cardoso, "Source separation using higher order moments," in *Proc. ICASSP*, Glasgow, U.K., May 1989, pp. 2109–2212.
[4] C. Jutten and J. Hérault, "Blind separation of sources, Part I: An adaptive algorithm based on a neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.
[5] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.
[6] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, 1995.
[7] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Process.*, vol. 45, pp. 59–83, 1995.
[8] J. Karhunen, E. Oja, L. Wang, R. Vigário, and J. Joutsensalo, "A class of neural networks for independant component analysis," *IEEE Trans. Neural Networks*, vol. 8, pp. 486–504, May 1997.
[9] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasimaximum likelihood approach," *IEEE Trans. Signal Processing*, vol. 45, pp. 1712–1725, July 1997.
[10] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Signal Processing*, vol. 10, pp. 626–634, May 1999.
[11] D. T. Pham, "Blind separation of instantaneous mixture of sources based on order statistics," *IEEE Trans. Signal Processing*, vol. 48, pp. 363–375, 2000.
[12] L. Tong and V. Soon, "Indeterminacy and identifiability of blind identification," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 499–509, May 1991.

[13] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlation," *Phys. Rev. Lett.*, vol. 72, pp. 3634–3636, 1994.
[14] A. Belouchrani, K. A. Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation technique based on second order statistics," *IEEE Trans. Signal Processing*, vol. 45, pp. 434–444, Feb. 1997.
[15] D. T. Pham, "Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion," *Signal Process.*, vol. 81, pp. 855–870, 2001.
[16] A. Ziehe and K. R. Müller, "Tdsep-An efficient algorithm for blind separation using time structure," in *Proc. Int. Conf. Artificial Neural Networks*, Skövde, Sweden, 1998, pp. 675–680.
[17] S. Degerine and R. Malki, "Second order blind separation of sources based on canonical partial innovations," *IEEE Trans. Signal Processing*, vol. 48, pp. 629–641, 2000.
[18] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.
[19] M. Gaeta and J.-L. Lacoume, "Source separation without a priori knowledge: The maximum likelihood solution," in *Proc. EUSIPCO*, vol. 2, Barcelona, Spain, Sept. 1990, pp. 621–624.
[20] A. Zaïdi, "Séparation aveugle d'un mélange instantané de sources autorégressives gaussiennes par la méthode du maximum de vraisemblance exact," Ph.D. dissertation (in French), Inst. Nat. Polytech. Grenoble, Grenoble, France, 2000.
[21] S. I. Amari, "Estimating functions of independent component analysis for temporally correlated signals," *Neural Comput.*, vol. 12, no. 9, pp. 2083–2107, 2000.
[22] D. T. Pham, "Mutual information approach to blind separation of stationary sources," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1935–1946, July 2002.
[23] ——, "Blind separation of instantaneous mixture of sources via an independant component analysis," *IEEE Trans. Signal Processing*, vol. 44, pp. 2768–2779, Nov. 1996.
[24] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 9, pp. 2009–2025, Oct. 1998.
[25] S. Hosseini, C. Jutten, and D. T. Pham, "Blind separation of temporally correlated sources using a quasi maximum likelihood approach," in *Proc. ICA*, San Diego, CA, Dec. 2001, pp. 586–590.
[26] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1972.
[27] D. T. Pham, "Estimation des fonctions score conditionnelles," Lab. Modélisation et Calcul, Grenoble, France, IMAG- C.N.R.S., 2002.
[28] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 3017–3030, Dec. 1996.



Shahram Hosseini was born in Shiraz, Iran, in 1968. He received the B.Sc. and M.Sc. degrees from Sharif University of Technology, Tehran, Iran, in 1991 and 1993, respectively, both in electrical engineering, and the Ph.D. degree from the Institut National Polytechnique (INPG), Grenoble, France, in 2000, in signal processing.

He was assistant professor of electrical engineering at INPG, from 2000 to 2001 and postdoctoral fellow with the Laboratoire des Images et des Signaux, Grenoble, from 2001 to 2002. He is currently an assistant professor with Paul Sabatier University, Toulouse, France. His research interests include artificial neural networks, blind separation of sources, and adaptive signal processing.

Christian Jutten (A'92) received the Ph.D. degree in 1981 and the Doc.Sci. degree in 1987 from the Institut National Polytechnique of Grenoble, Grenoble, France.

He was an associate professor with the Ecole Nationale Supérieure d'Electronique et de Radioélectricité of Grenoble from 1982 to 1989. He was visiting professor with the Swiss Federal Polytechnic Institute, Lausanne, Switzerland, in 1989, before becoming a Full Professor with the Sciences and Techniques Institute, Université Joseph Fourier, Grenoble. For 20 years, his research interests have been source separation and independent component analysis and learning in neural networks, including applications in signal processing (biomedical, seismic, speech) and data analysis. He is the author or coauthor of more than 30 papers in international journals and 70 communications in international conferences.

Dr. Jutten has been associate editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS from 1994 to 1995 and co-organizer, with Dr. J.-F. Cardoso and Prof. Ph. Loubaton, of the First International Conference on Blind Signal Separation and Independent Component Analysis (Aussois, France, January 1999). He is currently member of a technical committee of the IEEE Circuits and Systems Society on blind signal processing.

Dinh-Tuan Pham (M'88) was born in Hanoï, VietNam, on February 10, 1945. He graduated from the School of Applied Mathematics and Computer Science (ENSIMAG), Polytechnic Institute of Grenoble, Grenoble, France, in 1968 and received the Ph.D. degree in statistics in 1975 from the University of Grenoble.

He was a Postdoctoral Fellow with the Department of Statistics, University of California at Berkeley, from 1977 to 1978 and was a Visiting Professor with the Department of Mathematics, Indiana University, Bloomington, from 1979 to 1980. He is currently Director of Research at the French Centre National de Recherche Scientifique, Grenoble. His research interests include time series analysis, signal modeling, blind source separation, array processing, and biomedical signal processing.