BLIND SEPARATION OF INSTANTANEOUS MIXTURES OF NON STATIONARY SOURCES

Dinh-Tuan Pham

Laboratoire de Modélisation et Calcul
URA 397, CNRS/UJF/INPG
BP 53X, 38041 Grenoble cédex, France
Dinh-Tuan.Pham@imag.fr

Jean-François Cardoso

Centre National de la Recherche Scientifique
(C.N.R.S.), ENST-TSI
46 rue Barrault, 75634 Paris, France
cardoso@tsi.enst.fr

ABSTRACT

Most ICA algorithms are based on a model of stationary sources. This paper considers exploiting the (possible) non-stationarity of the sources to achieve separation. We introduce two objective functions based on the likelihood and on mutual information in a simple Gaussian non stationary model and we show how they can be optimized, off-line or on-line, by simple yet remarkably efficient algorithms (one is based on a novel joint diagonalization procedure, the other on a Newton-like technique). The paper also includes (limited) numerical experiments and a discussion contrasting non-Gaussian and non-stationary models.

1. INTRODUCTION

The aim of this paper is to develop a blind source separation procedure adapted to source signals with time varying intensity (such as speech signals). For simplicity, we shall restrict ourselves to the simplest mixture model:

$$\mathbf{X}(t) = \mathbf{A}\mathbf{S}(t) \tag{1}$$

where $\mathbf{X}(t) = [X_1(t) \ \cdots \ X_K(t)]^{\mathrm{T}}$ is the vector of observations (at time t), $\mathbf{A}$ is a fixed unknown $K \times K$ invertible matrix and $\mathbf{S}(t) = [S_1(t) \ \cdots \ S_K(t)]^{\mathrm{T}}$ is the vector of source sequences and $^{\mathrm{T}}$ denotes the transpose. The goal is to reconstruct the sources $S_k(t)$ based *only on the assumption of their mutual independence.*

Most of the approaches to blind source separation are based (explicitly or not) on a model where, for each i, $\{S_i(t)\}$ is a sequence of independently and identically distributed (i.i.d) variables (see [3] for a review of this approach). In this case, the blind identification of $\mathbf{A}$ is possible only if at most one of the sources has a Gaussian (marginal) distribution. In contrast, if the source sequences are not i.i.d., it is possible to blindly identify $\mathbf{A}$ even for Gaussian processes. This is the case when each source sequence is a stationary (possibly Gaussian) process with non proportional spectra [10, 12, 2] and when sources are non stationary processes [11, 13, 7, 8, 4]. In this paper, we derive objective functions from a simple non stationary model and introduce algorithms for their optimization, leading to very efficient separation techniques for non stationary sources.

2. OBJECTIVE FUNCTIONS

Using a simple non stationary model, we derive in this section two objective functions based on the maximum likelihood and minimum mutual information principles. In order to exploit non stationarity, we shall make the simplest distributional assumptions compatible with it: the sources are temporally independent and are Gaussian with a time dependent variance. We must stress that this is only a *working assumption* in order to derive objective functions. By making the independence assumption, we simply have chosen not to exploit the time dependence of the source signals and by making the Gaussian assumption, we have chosen to base the our procedures on second order statistics only. However, our algorithms are applicable even for colored non Gaussian sources (see section 4 for instance).

2.1. Maximum likelihood

The maximum likelihood (ML) objective is more conveniently handled by considering the negative of the normalized log probability density of the data set $\mathbf{X}(1), \ldots, \mathbf{X}(T)$, which we denote by C_{ML}. Under the Gaussian temporally independent model:

$$C_{ML} = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2}\mathrm{tr}[\mathbf{\Sigma}^{-2}(t)\mathbf{A}^{-1}\mathbf{X}(t)\mathbf{X}(t)^{\mathrm{T}}\mathbf{A}^{-\mathrm{T}}] + \frac{1}{2}\log\det[2\pi\mathbf{\Sigma}^2(t)] + \log|\det\mathbf{A}| \tag{2}$$

where tr denotes the trace, $\mathbf{A}^{-\mathrm{T}}$ stands for $(\mathbf{A}^{-1})^{\mathrm{T}}$ (for short) and $\mathbf{\Sigma}^2(t)$ is the covariance matrix of $\mathbf{S}(t)$, which is diagonal with diagonal elements $\sigma_1^2(t), \ldots, \sigma_K^2(t)$.

The variation of C_{ML} with respect to $\mathbf{A}$ is better expressed by computing its relative gradient, that is, the $K \times K$ matrix denoted $\mathbf{G}$, such that $C_{ML}(\mathbf{A} + \mathbf{A}\mathcal{E}) = C_{ML}(\mathbf{A}) + \mathrm{tr}(\mathcal{E}^{\mathrm{T}}\mathbf{G}) + o(\|\mathcal{E}\|)$. One finds

$$\mathbf{G} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{\Sigma}^{-2}(t) \hat{\mathbf{S}}(t) \hat{\mathbf{S}}(t)^{\mathrm{T}} - \mathbf{I} \tag{3}$$

where $\hat{\mathbf{S}}(t) = \mathbf{A}^{-1}\mathbf{X}(t)$. The stationary points (with respect to variations of $\mathbf{A}$) of the likelihood are characterized by $\mathbf{G} = 0$. The off diagonal elements of this matrix equation are:

$$\frac{1}{T} \sum_{t=1}^{T} \hat{S}_i(t) \hat{S}_j(t) / \sigma_i^2(t) = 0 \quad (1 \leq i \neq j \leq K) \tag{4}$$

$\hat{S}_i(t)$ being the i-th component of $\hat{\mathbf{S}}(t)$. These equations express some form of non-correlation between the reconstructed sources. The diagonal conditions merely state that the normalized reconstructed sources $\hat{S}_i/\sigma_i$ must have unit sample variance, thus determining the "scale factor" in $\mathbf{A}$.

In most practical situations, the variance profiles $\sigma_i^2(t)$ are not known in advance and must also be estimated from the data. The standard ML approach is to postulate a parametric model for these profiles. In a blind context, however, a non parametric approach is to be preferred: we simply estimate $\sigma_i^2(t)$ as a smoothed version of $\hat{S}_i^2(t)$. Note that *the estimate of* $\sigma_i^2(t)$ *needs not be consistent* because the decorrelation condition $\mathrm{E}[S_i(t)S_j(t)/\sigma_i^2(t)] = 0$, for which (4) is an empirical version, holds for zero mean independent sources, even if $\sigma_i^2(\cdot)$ is not the true variance profile.

2.2. Block Gaussian likelihood

In this section, we consider a 'block Gaussian' model in which the interval $[0, T]$ may be divided into L consecutive subintervals $T_1, \ldots, T_L$ such that $\sigma_i^2(t) = \sigma_{i,l}^2$ for $t \in T_l$, for all $i = 1, \ldots, K$. Define the matrices

$$\mathbf{R}_l = \mathbf{A}\mathbf{\Sigma}_l^2\mathbf{A}^{\mathrm{T}}, \qquad \hat{\mathbf{R}}_l = \frac{1}{\#T_l} \sum_{t \in T_l} \mathbf{X}(t)\mathbf{X}(t)^T \tag{5}$$

where $\mathbf{\Sigma}_l^2$ is the diagonal matrix with diagonal elements $\sigma_{1,l}^2$ $\ldots$, $\sigma_{K,l}^2$ and $\#T_l$ denotes the number of elements of T_l. Then the normalized log likelihood (2) can be expressed as

$$C_{ML} = \frac{1}{2} \sum_{l=1}^{L} w_l D\{\hat{\mathbf{R}}_l | \mathbf{R}_l\} + \text{Constant} \tag{6}$$

where $w_l = \#T_l/T$ is the proportion of data points in the l-th subinterval and

$$D\{\mathbf{R}_a | \mathbf{R}_b\} = \mathrm{tr}(\mathbf{R}_b^{-1}\mathbf{R}_a) - \log\det(\mathbf{R}_b^{-1}\mathbf{R}_a) - K \tag{7}$$

denotes the Kullback-Leibler divergence between two zero mean K-variate normal densities with covariance matrices $\mathbf{R}_a$ and $\mathbf{R}_b$ respectively. It is known that $D\{\mathbf{R}_a|\mathbf{R}_b\} \geq 0$ with equality if and only if $\mathbf{R}_a = \mathbf{R}_b$ and thus is a legitimate measure of deviation between positive matrices. Further, for $\mathbf{R}_l$ of the form $\mathbf{A}\mathbf{\Sigma}_l^2\mathbf{A}^{\mathrm{T}}$, we have $D\{\hat{\mathbf{R}}_l|\mathbf{R}_l\} = D\{\mathbf{A}^{-1}\hat{\mathbf{R}}_l\mathbf{A}^{-\mathrm{T}}|\mathbf{\Sigma}_l^2\}$ and therefore

$$C_{ML} = \frac{1}{2} \sum_{l=1}^{L} w_l D\{\mathbf{A}^{-1}\hat{\mathbf{R}}_l\mathbf{A}^{-\mathrm{T}} | \mathbf{\Sigma}_l^2\} + \text{Constant}. \tag{8}$$

For any positive $\mathbf{R}$ and any positive diagonal $\mathbf{\Sigma}$, the divergence $D\{\mathbf{R}|\mathbf{\Sigma}\}$ can be decomposed as:

$$D\{\mathbf{R}|\mathbf{\Sigma}\} = D\{\mathbf{R}|\mathrm{diag}\mathbf{R}\} + D\{\mathrm{diag}\mathbf{R}|\mathbf{\Sigma}\} \tag{9}$$

where $\mathrm{diag}\mathbf{R}$ denotes the diagonal matrix with the same diagonal as $\mathbf{R}$. Let us then define

$$\mathrm{off}(\mathbf{R}) = D\{\mathbf{R}|\mathrm{diag}\mathbf{R}\}, \tag{10}$$

which measures deviation from diagonality since it is non negative and can be zero only if it argument is diagonal. Using (9), the likelihood criterion (8) is seen to be minimized for a fixed value of $\mathbf{A}$ when $\mathbf{\Sigma}_l^2 = \mathrm{diag}(\mathbf{A}^{-1}\hat{\mathbf{R}}_l\mathbf{A}^{-\mathrm{T}})$ and the attained minimum is

$$C_{ML}^{\star} = \sum_{l=1}^{L} w_l \, \mathrm{off}(\mathbf{A}^{-1}\hat{\mathbf{R}}_l\mathbf{A}^{-\mathrm{T}}) + \text{Constant}. \tag{11}$$

It is very striking that the 'block-Gaussian' likelihood directly leads to an objective function which is a criterion of joint diagonalization. The idea of joint approximate diagonalization has already been used for source separation under different hypothesis: non Gaussian sources in [5], colored processes in [2]. In these contributions, however, the measure of joint diagonality was a simple quadratic criterion, not directly related to the likelihood objective and moreover is optimized under an orthogonality constraint which requires prior whitening of the observations.

2.3. Gaussian mutual information

We turn to a different objective: finding a transformation matrix $\mathbf{B}$ which minimizes of the mutual information between the random vectors

$$[(\mathbf{BX})_k(1) \;\; \cdots \;\; (\mathbf{BX})_k(T)]^{\mathrm{T}}, \quad k = 1, \ldots, K \tag{12}$$

Rather than trying to estimate the actual mutual information, we shall consider instead the Gaussian mutual information, defined in the same way as the ordinary mutual information but with respect to some hypothetical Gaussian random vectors which have the same covariance structure as the random vectors of interest. As we shall see, thanks to the non stationarity of the model, using the Gaussian mutual information still allows to achieve separation. Since the Kullback-Leibler divergence between the two Gaussian densities of zero mean and covariance matrices $\mathbf{P}$ and $\mathbf{Q}$ is $D\{\mathbf{P}|\mathbf{Q}\}$, the normalized Gaussian mutual information between the vectors (12) equals $\frac{1}{T}\sum_{t=1}^{T}\text{off}[\mathbf{B}\mathbf{R}(t)\mathbf{B}^T]$where $\mathbf{R}(t)$ denotes the covariance matrix of $\mathbf{X}(t)$. In practice, matrix $\mathbf{R}(t)$ is unknown; a sensible approach is to replace it by some non parametric kernel estimator:

$$\hat{\mathbf{R}}(t) = \frac{\sum_{\tau=1}^{T} k(\frac{t-\tau}{M})\mathbf{X}(\tau)\mathbf{X}(\tau)^{\mathrm{T}}}{\sum_{\tau=1}^{T} k(\frac{t-\tau}{M})}$$

where k is a positive kernel function and M is a window width parameter. The separation procedure then consists of minimizing $\frac{1}{T}\sum_{t=1}^{T}\text{off}[\mathbf{B}\hat{\mathbf{R}}(t)\mathbf{B}^{\mathrm{T}}]$ with respect to $\mathbf{B}$. But as $\hat{\mathbf{R}}(t)$ should vary slowly with t, one may approximate the above criterion by

$$C_{MI} = \frac{1}{L}\sum_{l=1}^{L}\text{off}[\mathbf{B}\hat{\mathbf{R}}(lT/L)\mathbf{B}^T] \qquad (13)$$

with L being some integer not exceeding T. The role of L is only to reduce the computation cost. There is little to gain by taking large L, since then the successive matrices $\hat{\mathbf{R}}(lT/L)$ would be very similar.

2.4. Discussion

Connections. It is not a coincidence that the above approaches lead to similar separating objectives. This is because the expectation of (2) is (up to a constant) a Kullback-Leibler divergence while the criterion (13) originates from a related Kullback-Leibler divergence. One can also compare these approaches on the basis of the corresponding estimating equations. The minima of C_{MI} are easily shown to be solution of

$$\frac{1}{L}\sum_{l=1}^{L}\frac{\widehat{S_iS_j}(lT/L)}{\widehat{S_iS_i}(lT/L)} = 0, \qquad 1 \le i \ne j \le K \qquad (14)$$

where, with $\hat{S}_i$ denoting the i-th component of $\hat{\mathbf{B}}\mathbf{X}$, we set:

$$\widehat{S_iS_j}(t) = \frac{\sum_{\tau=1}^{T} k(\frac{t-\tau}{M})\hat{S}_i(\tau)\hat{S}_j(\tau)}{\sum_{\tau=1}^{T} k(\frac{t-\tau}{M})}.$$

These equations are quite similar to (4), except that $S_i(t)S_j(t)$ and $\sigma_i^2(t)$ are replaced by local averages of S_iS_j and of S_i^2 around the time point t and that the time average in (4) is sparser, using a time step of T/L instead of 1.

Super efficiency. An interesting feature in the noise free non stationary setting is that there is room for ‘super efficiency’, that is, for estimating the mixing matrix with an error which decreases faster than $1/\sqrt{T}$. Assume that the i-th source is silent over a given interval $\mathcal{T}$ and the other sources are not, then there exists a vector $\mathbf{b}_i$ such that $\mathbf{b}_i^{\mathrm{T}}\mathbf{X}(t) = 0$ for all $t \in \mathcal{T}$. Since this vector must be orthogonal to all columns of $\mathbf{A}$ but the i-th column, it is proportional to the i-th row of $\mathbf{A}^{-1}$. Therefore this row can be determined without error from a finite number of samples. Summarizing the data in the interval $\mathcal{T}$ by the sample covariance matrix $\hat{\mathbf{R}}_{\mathcal{T}}$ preserves the possibility of error free because the matrix $\hat{\mathbf{R}}_{\mathcal{T}}$ although subjected to estimation errors always has its null space spanned by $\mathbf{b}_i$ and this is all that matters for finding the i-th row of $\mathbf{A}^{-1}$ without error.

In practice, a situation allowing super efficiency is unlikely to occur (for one thing, some noise is always present). But it is a guarantee of statistical effectiveness that a criterion yields super efficient estimates whenever such a possibility exists. This is the case of criterion (11).

3. ALGORITHMS

3.1. Block algorithm

The block Gaussian likelihood criterion (11) can be efficiently minimized thanks a novel joint approximate diagonalization algorithm which is now briefly described (see [9] for more details). Given positive matrices $\hat{\mathbf{R}}_1, \ldots, \hat{\mathbf{R}}_L$ and a set $w_1, \ldots, w_L$ of positive weights, it computes a matrix $\mathbf{B}$ minimizing $\sum_{l=1}^{L} w_l\text{off}(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^{\mathrm{T}})$. It works similarly to the classic Jacobi method by making successive transformations on each pair of rows of $\mathbf{B}$, but the transformations here are *not* constrained to be orthogonal. Explicitly, let $\mathbf{B}_{i\cdot}$ and $\mathbf{B}_{j\cdot}$ be any two distinct rows of $\mathbf{B}$. The algorithm changes $\mathbf{B}$ into a new matrix with these rows given by

$$\begin{bmatrix}\mathbf{B}_{i\cdot}\\ \mathbf{B}_{j\cdot}\end{bmatrix} \leftarrow \begin{bmatrix}\mathbf{B}_{i\cdot}\\ \mathbf{B}_{j\cdot}\end{bmatrix} - \mathbf{T}_{ij}\begin{bmatrix}\mathbf{B}_{i\cdot}\\ \mathbf{B}_{j\cdot}\end{bmatrix}, \qquad (15)$$

the other rows being unchanged. The 2×2 matrix $\mathbf{T}_{ij}$ can be chosen such that the criterion is sufficiently decreased. The procedure is then repeated with another pair of rows. The processing of all the $K(K-1)/2$ pairs is called a *sweep*. The algorithm consists in repeated sweeps until convergence is reached. Matrix $\mathbf{T}_{ij}$ in (15) is computed as

$$\mathbf{T}_{ij} = \frac{2}{1+\sqrt{1-4h_{ij}h_{ji}}}\begin{bmatrix}0 & h_{ij}\\ h_{ji} & 0\end{bmatrix} \qquad (16)$$

with the following definitions (which assume $\sum_{l=1}^{L} w_l = 1$; otherwise the weights must be renormalized)

$$g_{ij} = \sum_{l=1}^{L} w_l\frac{(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^{\mathrm{T}})_{ij}}{(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^{\mathrm{T}})_{ii}}, \quad \omega_{ij} = \sum_{l=1}^{L} w_l\frac{(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^{\mathrm{T}})_{jj}}{(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^{\mathrm{T}})_{ii}},$$

$$\begin{bmatrix}h_{ij}\\ h_{ji}\end{bmatrix} = \begin{bmatrix}\omega_{ij} & 1\\ 1 & \omega_{ji}\end{bmatrix}^{-1}\begin{bmatrix}g_{ij}\\ g_{ji}\end{bmatrix}. \qquad (17)$$

3.2. On-line algorithms

a. Simple stochastic gradient. This algorithm is based on the relative gradient (3) of the likelihood criterion (2). The separating matrix $\mathbf{B}(t)$ is updated upon reception of a new sample $\mathbf{X}(t)$ according to

$$\hat{\mathbf{B}}(t+1) = \hat{\mathbf{B}}(t) - \lambda \mathbf{G}(t)\hat{\mathbf{B}}(t) \tag{18}$$

where λ is a small positive constant and

$$\mathbf{G}(t) = \hat{\mathbf{\Sigma}}^{-2}(t)\hat{\mathbf{S}}(t)\hat{\mathbf{S}}(t)^{\mathrm{T}} - \mathbf{I} \quad \text{with} \quad \hat{\mathbf{S}}(t) = \mathbf{B}(t)\mathbf{X}(t).$$

Here $\hat{\mathbf{\Sigma}}^2(t)$ is the diagonal matrix with diagonal element $\hat{\sigma}_1^2(t), \ldots, \hat{\sigma}_K^2(t)]$ being some parametric estimates of $\sigma_k^2(t)$. For instance

$$\hat{\sigma}_k^2(t) = \hat{\sigma}_k^2(t-1) + \rho\,[\hat{S}_k^2(t) - \hat{\sigma}_k^2(t-1)] \tag{19}$$

where ρ is a small positive learning step, which must be significantly greater than λ since the estimated separating matrix $\hat{\mathbf{B}}$ should be nearly constant in a large range of time in which the source variances can vary significantly. This is the most straightforward algorithm but it can be significantly enhanced as follows.

b. On-line Newton-like technique. Consider an exponentially weighted relative gradient matrix $\bar{\mathbf{G}}_t(\mathbf{B})$ similar to (3):

$$\bar{\mathbf{G}}_t(\mathbf{B}) = \sum_{\tau \le t} \lambda(1-\lambda)^{t-\tau}\mathbf{\Sigma}^{-2}(\tau)\mathbf{B}\mathbf{X}(\tau)\mathbf{X}(\tau)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} - \mathbf{I} \tag{20}$$

computed at time t based on the past samples. As before, λ is a small positive parameter and $\mathbf{\Sigma}^2(\tau)$ is the diagonal matrix with diagonal elements $\sigma_1^2(\tau), \ldots, \sigma_K^2(\tau)$, assumed known for the moment. Our plan is to solve $\bar{\mathbf{G}}_t(\hat{\mathbf{B}}(t)) = 0$ assuming that this equation has been solved at time $t-1$. Similarly to (18), we write the solution at time t as a relative variation

$$\hat{\mathbf{B}}(t) = \hat{\mathbf{B}}(t-1) - \lambda \mathbf{H}(t)\hat{\mathbf{B}}(t-1). \tag{21}$$

A first order expansion shows that, if $\bar{\mathbf{G}}_{t-1}(\mathbf{B}) = 0$, then

$$\begin{aligned}\bar{\mathbf{G}}_t(\mathbf{B} - \lambda\mathbf{H}\mathbf{B}) \approx \lambda[\mathbf{\Sigma}^{-2}(t)\mathbf{B}\mathbf{X}(t)\mathbf{X}(t)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} - \mathbf{I}] - \lambda\mathbf{H}^{\mathrm{T}} \\ -\lambda\sum_{\tau\le t}\lambda(1-\lambda)^{t-\tau}\mathbf{\Sigma}^{-2}(\tau)\mathbf{H}(t)\mathbf{\Sigma}^2(\tau),\end{aligned}$$

where we drop the terms of order λ^2 and we approximate $\mathbf{B}\mathbf{X}(\tau)\mathbf{X}(\tau)^{\mathrm{T}}\mathbf{B}^T$ by $\mathbf{\Sigma}^2(\tau)$. With this expansion, the off diagonal term of the matrix equation $\bar{\mathbf{G}}_t[\mathbf{B}(t)] = 0$ yields

$$h_{ji} + h_{ij}\sum_{\tau\le t}\lambda(1-\lambda)^{t-\tau}\frac{\sigma_j^2(\tau)}{\sigma_i^2(\tau)} = \frac{\hat{S}_i(t)\hat{S}_j(t)}{\sigma_i^2(t)}, \tag{22}$$

for $1 \le i \ne j \le K$. Here, h_{ij} denotes the (i,j) entry of $\mathbf{H}(t)$ and we have set $\hat{S}_i(t) = [\mathbf{B}(t-1)\mathbf{X}(t)]_i$. We do not consider the equations for $i = j$: they only control the scales of the recovered sources, but such a control is not required. Using an on-line estimator for $\sigma_k^2(t)$, we obtain this algorithm:

1. Compute $\hat{\mathbf{S}}(t) = \mathbf{B}(t-1)\mathbf{X}(t)$, update $\hat{\sigma}_k^2(t)$ by (19) and $\hat{\omega}_{ij}(t)$ by

$$\hat{\omega}_{ij}(t) = \hat{\omega}_{ij}(t-1) + \lambda[\hat{\sigma}_j^2(t)/\hat{\sigma}_i^2(t) - \hat{\omega}_{ij}(t-1)]$$

2. Update $\hat{\mathbf{B}}(t)$ according to (21) where the diagonal of matrix $\mathbf{H}(t)$ is set to zero and its off diagonal elements are the solutions of (22) *i.e.*

$$\begin{bmatrix} h_{ij}(t) \\ h_{ji}(t) \end{bmatrix} = \begin{bmatrix} \hat{\omega}_{ij}(t) & 1 \\ 1 & \hat{\omega}_{ji}(t) \end{bmatrix}^{-1} \begin{bmatrix} \hat{S}_i(t)\hat{S}_j(t)/\hat{\sigma}_i^2(t) \\ \hat{S}_j(t)\hat{S}_i(t)/\hat{\sigma}_j^2(t) \end{bmatrix} \tag{23}$$

As before, the parameter λ, should be much smaller than ρ.

c. On-line versions of batch algorithms. The block Gaussian approach can be easily turned into a block on-line algorithm. The data stream is subdivided into data blocks of a given length, m say. For the l-th data block, one computes the sample covariance matrix $\hat{\mathbf{R}}_l$ similarly to (5). The L most recent covariance matrices are kept in memory and, after block l has become available, one performs the joint approximate diagonalization of the matrices $\hat{\mathbf{R}}_l, \ldots, \hat{\mathbf{R}}_{l+1-L}$ to obtain a separating matrix. This approach may seem computationally demanding but it is not the case because, in the on line context, it is sensible to perform only a *single* sweep of the joint diagonalization algorithm after a new data block is received.

Likewise, the Gaussian mutual information approach of section 2.3 gives rise to a similar and somewhat more flexible on-line algorithm. The matrices $\hat{\mathbf{R}}_l$ can now be evaluated at any time point as a local average. This is best done by applying a low-pass filter to the matrix sequence $\mathbf{X}(t)\mathbf{X}(t)^{\mathrm{T}}$ which outputs *positive matrices*, such as the exponential filter. The separating matrix $\mathbf{B}(t)$ is then obtained by jointly approximately diagonalizing the matrices $\hat{\mathbf{R}}(t)$, $\hat{\mathbf{R}}(t-m), \ldots, \hat{\mathbf{R}}(t+m-mL)$. Here the role of m is to reduce the number of matrices to be diagonalized. As before, only one sweep of the joint approximate diagonalization algorithm is performed.

4. NUMERICAL EXPERIMENTS

On-line algorithms. We illustrate the improved behavior of the Newton-like algorithm over the standard relative gradient approach. We use synthetic source signals: $S_i(t) = a_i(t)n_i(t)$ where $n_i(\cdot)$ is a Gaussian i.i.d. sequence and

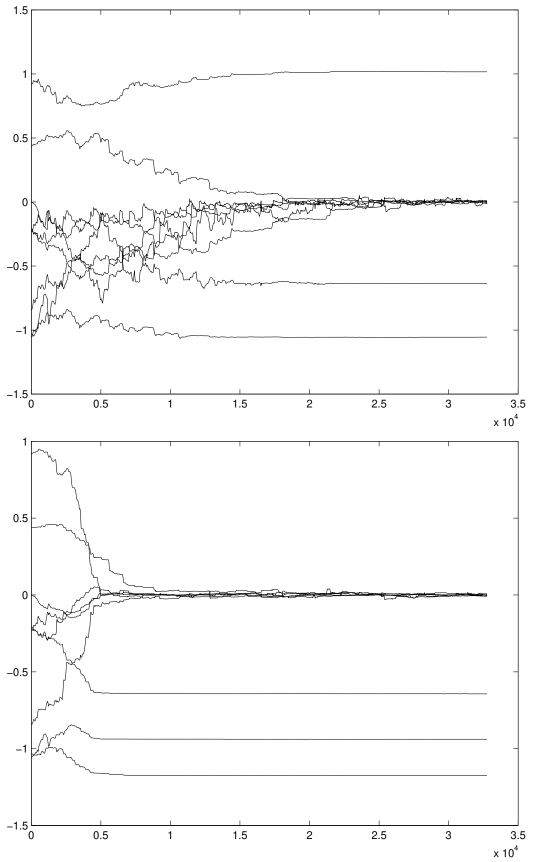

Figure 1: Convergence of the 9 coefficients of the global system $\mathbf{B}(t)\mathbf{A}$ for a $K = 3$ source case. Top: the 'regular' relative gradient technique. Bottom: the Newton-like technique.

$a_i(t)$ is a 'slowly varying' amplitude. These signals are mixed by a 3×3 matrix. Figure 1 shows the convergence of the 9 coefficients $(\mathbf{B}(t)\mathbf{A})_{ij}$ of the global system: the top panel is for the 'regular' relative gradient algorithm (18) with the diagonal of the relative gradient $\mathbf{G}(t)$ set to 0; the bottom panel is for the Newton-like algorithm (23). We have used the same signals, the same parameters ($\rho = 10^{-2}$ and $\lambda = \rho/20$) and the same starting point. The significantly faster convergence of the Newton-like algorithm is clearly visible.

Block on-line algorithms. Figure 2 shows the online version of the joint diagonalization algorithm separating a synthetic mixture of 3 speech waveforms (we use a block length of $m = 320$ samples (40 ms) and $L = 12$ matrices to be jointly diagonalized). The 9 coefficients of the global system $\mathbf{BA}$ are displayed versus the number of blocks. The convergence is reached after about 11 blocks, that is even before the memory is full.

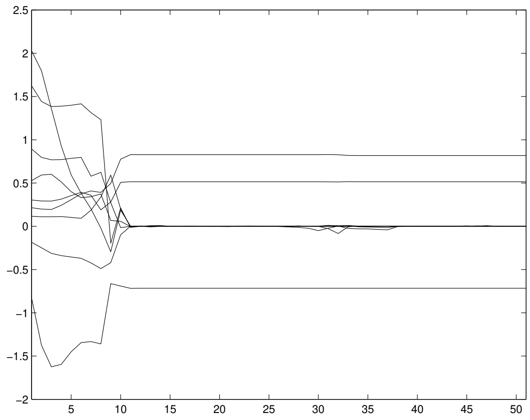

Figure 2: Convergence for the on-line joint diagonalizer.

5. DISCUSSION

Connections. The efficient approaches considered in this paper —the joint diagonalization algorithm of section 3.1 and the Newton-like algorithm of section 3.2— bear some resemblance: in both cases, a key step is the transformation of the gradient g_{ij} into a 'rectified gradient' h_{ij} (compare eq. (17) and (23) and the related updating rules). Here, the underlying mechanism can be recognized as the classic Newton technique in which the gradient is left multiplied by the inverse of the Hessian for it to point in the best (in a certain sense) direction. It is likely that the 'natural gradient' approach of Amari [1] would result in similar algorithms. We note however that the on-line algorithm takes its particular simple form thanks to an approximation which is only valid when the model holds even though the algorithm still behaves well if this is not the case.

About non stationarity. Another line of comments regards the notion of non stationarity used in this paper. In essence, the source properties which make the algorithms work are source independence and slowly varying variance profiles. In full rigor, the latest is not related to the well defined notion of stationarity. Indeed, considered the case where the i-th source signal is $S_i(t) = a_i(t)n_i(t)$ with $n_i(t)$ is an i.i.d. sequence and $a_i(t)$ a slowly varying *stationary* process. Strictly speaking, $S_i(t)$ is stationary even though visual inspection of a T sample realization $S_i(1), \ldots, S_i(T)$ shows a waveform which is 'psychologically non stationary'. Linear mixtures of such stationary sequences can actually be successfully separated by our algorithms. Conversely, it is easy to construct non stationary source processes with constant variance, which would defeat our algorithms. In summary, it would be more accurate to describe our algorithms as applying to independent sources with 'slow' amplitude modulation.

Non stationarity and non Gaussianity. A final comment regards a connection between non stationarity and non Gaussianity. For simplicity, consider again the model $S_i(t) = a_i(t)n_i(t)$ above. If the time index is ignored, as is done in 'classic' non Gaussian source separation techniques, then T successive samples of $S_i(t)$ are (implicitly) considered as T realizations of an i.i.d. sequence and the sample distribution will be strongly non Gaussian (even if $n_i(t)$ are Gaussian) if the amplitude $a_i(t)$ varies significantly over $[1, T]$.

Another direct connection to non Gaussian technique is as follows. If we do not assume that the variance profiles are smoothly varying, then each variance $\sigma_i^2(t)$ is a free parameter. In this case, the ML estimator of $\sigma_i^2(t)$ would be $\hat{S}_i^2(t)$ which is certainly not very engaging. A Bayesian estimate can be obtained by assigning a prior distribution to $\sigma_i(t)$ and estimating it as the mode or as the mean of its posterior distribution given $\hat{S}_i(t)$. Using an inverse gamma prior, the regularized variance estimate simply is

$$\hat{\sigma}_i^2(t) = \frac{\hat{S}_i^2(t) + n_0\sigma_0^2}{1 + n_0}$$

where n_0 and σ_0^2 are free hyper parameters (to be interpreted as n_0 extra data points with sample variance σ_0^2 [6]). In this case, the estimating equation $\mathbf{G} = 0$ becomes

$$\frac{1}{T}\sum_t \psi[\hat{S}_i(t)]\hat{S}_j(t) - \delta_{ij} = 0 \qquad (24)$$

where ψ is the non-linear function $\psi(y) = \frac{y(1+n_0)}{y^2+n_0\sigma_0^2}$. In other words, we end up with the exact same type of estimating equations that is obtained in i.i.d. (stationary) non Gaussian modeling! The simplest choices: $n_0 = 1$ and $\sigma_0 = 1$ yield $\psi(y) = \frac{2y}{y^2+1}$, which is minus the log derivative of the Cauchy density. In other words, solving eq. (24) amounts to using a model of i.i.d Cauchy sources.

Relation to previous works. Matsuoka *et al.* [7] consider an objective function which is essentially identical to ours but do not relate it to mutual information or maximum likelihood and do not propose an efficient algorithm for its optimization. Souloumiac [11] and Tsatsanis [13] consider the case of two distinct stationary regimes and actually perform a joint diagonalization of the two corresponding covariance matrices. The diagonalization is exact but this approach is limited to a very simple non stationary scenario.

Conclusions. Mixtures of independent sources can be separated by exploiting their non stationarity. We have presented criteria and algorithms for this task which are efficient numerically (simple implementations, fast convergence) as well as statistically (potential super-efficiency). Future investigation will address the issue of jointly exploiting both non stationarity and non Gaussianity and will include the study of the asymptotic performance.

6. REFERENCES

[1] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[2] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and Éric Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Sig. Proc.*, 45(2):434–44, Feb. 1997.

[3] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE. Special issue on blind identification and estimation*, 9(10):2009–2025, Oct. 1998.

[4] J.-F. Cardoso. Séparation de sources non stationnaires. In *Proc. GRETSI, Vannes, France*, pages 741–744, 1999.

[5] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, Dec. 1993.

[6] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall, 1995.

[7] K. Matsuoka, M. Ohya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural networks*, 8(3):411–419, 1995.

[8] J. Ngo and N. Bhadkamkar. Adaptive blind separation of audio sources by a physically compact device using second-order statistics. In *Proc. ICA'99*, pages 257–260, Aussois, France, January 11–15, 1999.

[9] D. Pham. Joint approximate diagonalization of positive definite Hermitian matrices. Technical report LMC/IMAG, http://www-lmc.imag.fr/lmc-sms/Dinh-Tuan.Pham/jadiag/jadiag.ps.gz, Apr. 1999.

[10] D.-T. Pham and P. Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Tr. SP*, 45(7):1712–1725, July 1997.

[11] A. Souloumiac. Blind source detection and separation using second order nonstationarity. In *Proc. ICASSP*, pages 1912–1915, 1995.

[12] L. Tong, V. Soon, Y. Huang, and R. Liu. AMUSE: a new blind identification algorithm. In *Proc. ISCAS*, 1990.

[13] M. K. Tsatsanis and C. Kweon. Source separation using second order statistics: Identifiability conditions and algorithms. In *Proc. 32nd Asilomar Conf. on Signals, Systems, and Computers*, pages 1574–1578. IEEE, Nov. 1998.