

Blind Separation of Instantaneous Mixtures of Non Stationary Sources

Dinh-Tuan Pham, Member IEEE
Jean-François Cardoso, Member IEEE

To appear in IEEE Transactions on Signal Processing (accepted Oct. 2000).

D.T. Pham is with the Laboratory of Modeling and Computation, IMAG - C.N.R.S., B.P. 53X, 38041 Grenoble cedex 9, France. E-mail: Dinh-Tuan.Pham@imag.fr

J.F. Cardoso is with the Centre National de la Recherche Scientifique (C.N.R.S.), ENST-TSI, 46 rue Barrault, 75634 Paris, France E-mail: cardoso@tsi.enst.fr

DRAFT

Abstract

Most source separation algorithms are based on a model of stationary sources. However, it is a simple matter to take advantage of possible non-stationarities of the sources to achieve separation. This paper develops novel approaches in this direction, based on the principles of maximum likelihood and minimum mutual information. These principles are exploited by efficient algorithms in both the off-line case (via a new joint diagonalization procedure) and in the on-line case (via a Newton-like procedure). Some experiments are presented showing the good performance of our algorithms and evidencing an interesting feature of our methods: their ability to achieve a kind of super-efficiency. The paper concludes with a discussion contrasting separating methods for non-Gaussian and non-stationary models and emphasizing that, as a matter of fact, ‘what makes the algorithms work’ is —strictly speaking— not the non stationarity itself but rather the property that each realization of the source signals has a time-varying envelope.

Keywords

Independence. Joint approximate diagonalization. Kullback-Leibler divergence. Likelihood. Mutual information. Non stationarity. Separation of sources.

I. INTRODUCTION

This paper is concerned with the problem of blind source separation. In its simplest form, the underlying model is that of a sequence $\{\mathbf{X}(t)\}$ of K -dimensional samples modeled as a linear mixture of K sequences of ‘source signals’:

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

where \mathbf{A} is a fixed unknown $K \times K$ invertible matrix, $\mathbf{X}(t) = [X_1(t) \ \cdots \ X_K(t)]^T$ is the vector of observations, $\mathbf{S}(t) = [S_1(t) \ \cdots \ S_K(t)]^T$ is the vector of source sequences and the notation T denotes transposition. The objective is to reconstruct the sources $S_k(t)$ from the observations, exploiting only the assumption of mutual independence between the sources without relying on any precise knowledge about their distribution.¹

This model has recently received a lot of attention, in particular because it can be used to process data from multi-sensor measurements *without* requiring that the underlying physical phenomena be modeled accurately; actually this approach requires no modeling at all for the mixing part of the model. This is compensated by a strong assumption about the other part of the model: that the source sequences are mutually independent. This assumption leaves a wide range of options for modeling each individual source sequence; in this paper, we consider non stationary Gaussian source sequences.

Most of the approaches to blind source separation are based (explicitly or not) on a model where, for each i , the sequence $\{S_i(t)\}$ is a sequence of independently and identically distributed (i.i.d.) variables (see [1] for a review of this approach). In this case, blind identification of \mathbf{A} is impossible if the sources are normally distributed [2]. In contrast, if the source sequences are not i.i.d., it is possible to blindly identify \mathbf{A} even for Gaussian processes. For instance, blind identification is possible if the source processes are stationary processes with different spectra (see *e.g.* [3], [4], [5]). This paper considers the case when the second ‘i’ of ‘i.i.d.’ is failing, that is, the non stationary case. Previous contributions [6], [7], [8], [9], [10] to the non stationary case (and how we improve on them) are discussed in sections IV and V where the connections between non stationarity and non Gaussianity are also addressed. Intuitively, non stationarity allows blind identification in the Gaussian case because, while a single covariance matrix does not give enough constraints to uniquely determine \mathbf{A} , a *collection* of several covariance matrices estimated over different time periods does determine \mathbf{A} provided the source distributions have changed enough over the whole observation period.

¹As is customary, the sources are assumed to have zero mean, as it is often the case in practice.

Our focus being the exploitation of non stationarity, we shall make the simplest possible distributional assumptions compatible with it. The model under consideration assumes that the sources are temporally independent, that is, $S_i(t)$ is independent from $S_i(t')$ for $t \neq t'$. We must stress that this is only a *working assumption*, that is, it is used to build a statistical model simple enough to yield simple algorithms and rich enough to capture the non stationarity of the sources; the resulting algorithms in fact still work for a large class of colored (or temporally correlated) source sequences. By making this working assumption, we simply have chosen not to exploit the time dependence of the source signals; it does not imply that the source signals should have no time dependence to be separable with the proposed techniques (as will be seen in the experimental section). Likewise, we shall also make the working assumption that the sources are Gaussian. Again, the algorithms obtained via this simplifying assumption are in fact applicable to non Gaussian signals: the consequence of using a Gaussian model is that the resulting techniques are based on second order statistics only. One of the objectives of this paper is to devise source separation techniques exploiting second order statistics of non stationary observations. Even though we are looking for blind identifiability through non stationarity rather than through non Gaussianity, there is—at least at the algorithmic level—an interesting connection between these two aspects (see section V).

The paper is organized as follows. In section II, we investigate the structure of the likelihood under our working assumptions and we also consider a mutual information criterion. We shall see that these two points of view lead to very similar objective functions. In section III, we describe several techniques for the optimization of these objective functions which are illustrated by numerical experiments in section IV. A final section discusses in some detail our findings, the applicability of the algorithms and connections to other approaches to source separation.

II. OBJECTIVE FUNCTIONS

In the source separation problem, two principles—maximum likelihood and minimum mutual information—provide foundations for deriving objective functions. We investigate their specific form under our working assumptions of non stationary Gaussian sources.

A. Maximum likelihood

We follow the ‘quasi maximum likelihood’ approach of Pham and Garat [3]. The maximum likelihood objective is more conveniently handled by considering the negative of the normalized log probability of the data set (here ‘normalized’ refers to the division by the data length), which we denote by C_{ML} and refer to as the ‘likelihood criterion’ in the sequel. For a given batch of T data points, it is a function of a $K \times K$ matrix parameter of interest (the ‘mixing matrix’ \mathbf{A}) and of KT nuisance parameters (the variances of each source at each t).

In the Gaussian model, the log probability density of a source vector $\mathbf{S}(t)$ is

$$-\frac{1}{2} \sum_{k=1}^K \left\{ \frac{S_k^2(t)}{\sigma_k^2(t)} + \log[2\pi\sigma_k^2(t)] \right\} = -\frac{1}{2} \text{tr}[\boldsymbol{\Sigma}^{-2}(t)\mathbf{S}(t)\mathbf{S}(t)^T] - \frac{1}{2} \log \det[2\pi\boldsymbol{\Sigma}^2(t)] \quad (2)$$

where tr denotes the trace and where $\boldsymbol{\Sigma}^2(t)$ denotes the covariance matrix of $\mathbf{S}(t)$:

$$\boldsymbol{\Sigma}^2(t) = \text{diag}[\sigma_1^2(t), \dots, \sigma_K^2(t)],$$

and $\text{diag}(\cdot, \dots, \cdot)$ builds a diagonal matrix from its arguments. Since the observation vector is a linear invertible transformation $\mathbf{X} = \mathbf{A}\mathbf{S}$ of the source vector, its probability density $f_X(\mathbf{X})$ is simply related

to the density $f_S(\mathbf{S})$ of \mathbf{S} by $f_X(\mathbf{X}) = |\det \mathbf{A}^{-1}| f_S(\mathbf{A}^{-1}\mathbf{X})$. Therefore, using (2) the likelihood criterion C_{ML} is:

$$C_{ML} = \frac{1}{T} \sum_{t=1}^T \frac{1}{2} (\text{tr}(\boldsymbol{\Sigma}^{-2}(t) \mathbf{A}^{-1} \mathbf{X}(t) \mathbf{X}(t)^T \mathbf{A}^{-T}) + \log \det(2\pi \boldsymbol{\Sigma}^2(t))) + \log |\det \mathbf{A}| \quad (3)$$

where we have written \mathbf{A}^{-T} for $(\mathbf{A}^{-1})^T$ for short.

The variation of the likelihood criterion with \mathbf{A} is better expressed by considering *relative* variations: it is defined as the $K \times K$ matrix \mathbf{G} such that

$$C_{ML}(\mathbf{A} - \mathbf{A}\boldsymbol{\mathcal{E}}) = C_{ML}(\mathbf{A}) + \text{tr}(\boldsymbol{\mathcal{E}}^T \mathbf{G}) + o(\|\boldsymbol{\mathcal{E}}\|) \quad (4)$$

for any $K \times K$ matrix $\boldsymbol{\mathcal{E}}$.² The relative gradient matrix \mathbf{G} is readily obtained by using the expansions $(I - \boldsymbol{\mathcal{E}})^{-1} = I + \boldsymbol{\mathcal{E}} + o(\|\boldsymbol{\mathcal{E}}\|)$ and $\log |\det(I + \boldsymbol{\mathcal{E}})| = \text{tr}(\boldsymbol{\mathcal{E}}) + o(\|\boldsymbol{\mathcal{E}}\|)$ and collecting all the first order terms. One finds

$$\mathbf{G} = \frac{1}{T} \sum_{t=1}^T \boldsymbol{\Sigma}^{-2}(t) \hat{\mathbf{S}}(t) \hat{\mathbf{S}}(t)^T - \mathbf{I} \quad \text{or} \quad \mathbf{G}_{ij} = \frac{1}{T} \sum_{t=1}^T \frac{\hat{S}_i(t) \hat{S}_j(t)}{\sigma_i^2(t)} - \delta_{ij} \quad (5)$$

where $\hat{\mathbf{S}}(t)$ denotes the estimate of $\mathbf{S}(t)$ for a given value of \mathbf{A} , that is, $\hat{\mathbf{S}}(t) = \mathbf{A}^{-1} \mathbf{X}(t)$ and \hat{S}_i is the i -th component of $\hat{\mathbf{S}} = \mathbf{A}^{-1} \mathbf{X}$ and represents the reconstructed i -th source for a given value of the parameter \mathbf{A} .

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

$$\frac{1}{T} \sum_{t=1}^T \hat{S}_i(t) \hat{S}_j(t) / \sigma_i^2(t) = 0 \quad (1 \leq i \neq j \leq K) \quad (6)$$

which expresses some form of non-correlation between the reconstructed sources. The diagonal conditions $\mathbf{G}_{ii} = 0$ for $1 \leq i \leq K$ merely state that the normalized reconstructed sources \hat{S}_i / σ_i must have unit sample variance, hence they determine 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. It is clear however that these profiles cannot be reliably estimated without some kind of prior assumptions since there are as many values of $\sigma_i^2(t)$ as data points. The standard parametric ML approach would be to postulate a parametric model, *i.e.* $\sigma_i^2(t) = f(t; \theta_i)$ where $f(t; \theta_i)$ is a smooth function of t depending on a vector of parameters θ_i , to be estimated by solving:

$$\frac{\partial C_{ML}}{\partial \theta_i} = \frac{1}{T} \sum_{t=1}^T \frac{\hat{S}_i(t)^2 - f(t; \theta_i)}{f^2(t; \theta_i)} \cdot \frac{\partial f(t; \theta_i)}{\partial \theta_i} = 0, \quad (i = 1, \dots, K). \quad (7)$$

The specific form of the estimate would depend on the particular model $f(t; \theta)$ but there is no point giving more details here since we shall *not* try to solve the estimating equations (7). Rather, we note that these equations suggest —not surprisingly— that the estimate of the variance profile $\sigma_i^2(t)$ should depend only on the sequence $\hat{S}_i^2(t)$. This motivates (if necessary) a more straightforward *non parametric* approach by which $\sigma_i^2(t)$ is simply estimated as a smoothed version of $\hat{S}_i^2(t)$. At this point, it is worth stressing that this estimate needs not be consistent in order to get a consistent estimate of \mathbf{A} . Essentially, this is because the decorrelation condition $\text{E}[S_i(t) S_j(t) / \sigma_i^2(t)] = 0$ (of which eq. (6) is an empirical version) holds for zero-mean independent sources even if $\sigma_i^2(t)$ is not the true variance of $S_i(t)$. For this reason, it is only needed to obtain rough estimates of the variances (this is amply illustrated in the experiments of section IV).

²The choice of a minus sign in the left-hand side of definition (4) is for consistency with previous works which define relative gradient with respect to the *inverse* of \mathbf{A} . See, *e.g.* [11].

B. Block Gaussian likelihood

In this section, we consider a ‘block Gaussian’ model where the variance profiles are modeled as being constant over subintervals. Thanks to this specific assumption, the likelihood can be directly connected to a joint diagonalization criterion for which an efficient optimization exists (see below at section III-A).

Specifically, the interval $[0, T]$ is divided into L consecutive subintervals T_1, \dots, T_L and the model is that $\sigma_i^2(t) = \sigma_{i,l}^2$ for $t \in T_l$, for all $i = 1, \dots, K$. Defining the matrices

$$\mathbf{\Sigma}_l^2 = \text{diag}(\sigma_{1,l}^2, \dots, \sigma_{K,l}^2), \quad \mathbf{R}_l = \mathbf{A}\mathbf{\Sigma}_l^2\mathbf{A}^\text{T}, \quad \hat{\mathbf{R}}_l = \frac{1}{\#T_l} \sum_{t \in T_l} \mathbf{X}(t)\mathbf{X}(t)^\text{T} \quad (8)$$

where $\#T_l$ denotes the number of elements of T_l , the normalized log likelihood (3) can be rewritten as

$$C_{ML} = \frac{1}{2} \sum_{l=1}^L w_l [\text{tr}(\mathbf{R}_l^{-1} \hat{\mathbf{R}}_l) - \log \det(\mathbf{R}_l^{-1} \hat{\mathbf{R}}_l) - K] + \text{Constant} \quad (9)$$

where $w_l = \#T_l/T$ is the proportion of data points in the l -th subinterval. The constant term in (9) is equal to $\frac{1}{2} \sum_{l=1}^L w_l [\log \det(2\pi \hat{\mathbf{R}}_l) + K]$ as can be readily checked by direct substitution.

The Kullback-Leibler divergence $D\{\mathbf{R}_a | \mathbf{R}_b\}$ between two zero mean K -variate normal densities, with covariance matrices \mathbf{R}_a and \mathbf{R}_b respectively, is given by

$$D\{\mathbf{R}_a | \mathbf{R}_b\} = \frac{1}{2} [\text{tr}(\mathbf{R}_b^{-1} \mathbf{R}_a) - \log \det(\mathbf{R}_b^{-1} \mathbf{R}_a) - K]. \quad (10)$$

This divergence is a measure of deviation between probability distributions; our notation $D\{\mathbf{R}_a | \mathbf{R}_b\}$ specializes it as a measure of deviation between positive matrices. In particular, $D\{\mathbf{R}_a | \mathbf{R}_b\} \geq 0$ with equality only if $\mathbf{R}_a = \mathbf{R}_b$. In addition, for \mathbf{R}_l of the form $\mathbf{A}\mathbf{\Sigma}_l^2\mathbf{A}^\text{T}$, we have $D\{\hat{\mathbf{R}}_l | \mathbf{R}_l\} = D\{\mathbf{A}^{-1} \hat{\mathbf{R}}_l \mathbf{A}^{-\text{T}} | \mathbf{\Sigma}_l^2\}$. This is a consequence of the invariance of the Kullback-Leibler divergence under invertible transforms; in the Gaussian case, it can also be directly checked from (10). Thus, by (9)

$$C_{ML} = \sum_{l=1}^L w_l D\{\mathbf{A}^{-1} \hat{\mathbf{R}}_l \mathbf{A}^{-\text{T}} | \mathbf{\Sigma}_l^2\} + \text{Constant} \quad (11)$$

At this point, a key property of the Kullback divergence is that for a positive matrix \mathbf{R} and any positive diagonal matrix $\mathbf{\Sigma}$, the following (Pythagorean) decomposition holds [12]:

$$D\{\mathbf{R} | \mathbf{\Sigma}\} = D\{\mathbf{R} | \text{diag}(\mathbf{R})\} + D\{\text{diag}(\mathbf{R}) | \mathbf{\Sigma}\} \quad (12)$$

where $\text{diag}(\mathbf{R})$ denotes the diagonal matrix with the same diagonal as \mathbf{R} . Eq. (12) shows that the closest (in the $D\{\cdot | \cdot\}$ sense) diagonal matrix to \mathbf{R} is $\mathbf{\Sigma} = \text{diag}(\mathbf{R})$ since this choice cancels the (non negative) rightmost term in (12). Decomposition (12) and expression (11) make it trivial to minimize C_{ML} with respect to the nuisance parameters $\{\mathbf{\Sigma}_1, \dots, \mathbf{\Sigma}_L\}$ for a fixed value of \mathbf{A} : one must choose $\mathbf{\Sigma}_l^2 = \text{diag}(\mathbf{A}^{-1} \hat{\mathbf{R}}_l \mathbf{A}^{-\text{T}})$ and the attained minimum is

$$\sum_{l=1}^L w_l D\{\mathbf{A}^{-1} \hat{\mathbf{R}}_l \mathbf{A}^{-\text{T}} | \text{diag}(\mathbf{A}^{-1} \hat{\mathbf{R}}_l \mathbf{A}^{-\text{T}})\} + \text{Constant}.$$

For any positive matrix \mathbf{R} , the quantity

$$\text{off}(\mathbf{R}) = D\{\mathbf{R} | \text{diag} \mathbf{R}\}. \quad (13)$$

is a measure of deviation from diagonality: it is always non negative and equals zero only when \mathbf{R} is diagonal.³ Thus, optimizing the likelihood criterion with respect to the nuisance parameters leaves us with a ‘reduced’ objective denoted C_{ML}^* :

$$C_{ML}^* = \sum_{l=1}^L w_l \text{off}(\mathbf{A}^{-1} \hat{\mathbf{R}}_l \mathbf{A}^{-T}) \quad (14)$$

which depends only on the parameter of interest (we dropped the constant term in (11)).

It is very striking that the ‘block-Gaussian’ likelihood directly leads to an objective function which is a criterion of joint diagonality. The idea of joint approximate diagonalization has already been used for source separation under different hypothesis: non Gaussian sources in [14], colored processes in [5], using a different measure of diagonality, namely the quadratic criterion $\text{off}(\mathbf{R}) = \sum_{i \neq j} \mathbf{R}_{ij}^2$. More recently, Parra and Spence [10] has also proposed a quadratic criterion for non stationary sources. Such criteria are only approximately related to the likelihood [1] and, in addition, are to be optimized under a decorrelation constraint. As explained in [15], this constraint bounds the performance and forbids the ‘super efficiency’ effect discussed in section II-D.2 and illustrated in section IV.

C. Gaussian mutual information

Another approach to the blind source separation problem is to consider it in the light of an independent component analysis, which aims to find a transformation matrix \mathbf{B} such that the components of the transformed observation vectors $\mathbf{B}\mathbf{X}(t)$ are as independent as possible. A natural measure of independence is the mutual information. Let Y_1, \dots, Y_K be K random vectors, with joint density f_{Y_1, \dots, Y_K} and marginal densities f_{Y_1}, \dots, f_{Y_K} , the mutual information between them is defined as

$$I(Y_1, \dots, Y_K) = -\mathbb{E} \left[\log \frac{\prod_{k=1}^K f_{Y_k}(Y_k)}{f_{Y_1, \dots, Y_K}(Y_1, \dots, Y_K)} \right]$$

which is nothing but the Kullback-Leibler divergence between the density f_{Y_1, \dots, Y_K} and the product density $\prod_{k=1}^K f_{Y_k}$. The idea is thus to minimize the mutual information between the vectors $[(\mathbf{B}\mathbf{X})_k(1) \ \dots \ (\mathbf{B}\mathbf{X})_k(T)]^T$, $k = 1, \dots, K$. As the vector $\mathbf{X}(t)$ is temporally independent (by assumption), the (normalized) mutual information for the whole sequence $\mathbf{B}\mathbf{X}(t)$, $t = 1, \dots, T$ is

$$\frac{1}{T} \sum_{t=1}^T I[(\mathbf{B}\mathbf{X})_1(t), \dots, (\mathbf{B}\mathbf{X})_K(t)].$$

The above criterion requires the knowledge of the density of the sources, which is not practical. Therefore 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. Clearly, using the Gaussian mutual information would lead to a second order method, but this is still enough to achieve separation by exploiting the non stationarity. As already mentioned in previous section, the Kullback-Leibler divergence between the two K -variate Gaussian densities of zero mean and covariance matrices \mathbf{P} and \mathbf{Q} is $D\{\mathbf{P} | \mathbf{Q}\}$, hence the Gaussian mutual information between the random variables Y_1, \dots, Y_K is $D\{\mathbf{P} | \text{diag}\mathbf{P}\}$ where \mathbf{P} denotes the covariance matrix of the random vector $[Y_1 \ \dots \ Y_K]^T$. Therefore, denoting by $\mathbf{R}(t)$ the covariance matrix of $\mathbf{X}(t)$, the Gaussian mutual information criterion is $\frac{1}{T} \sum_{t=1}^T \text{off}[\mathbf{B}\mathbf{R}(t)\mathbf{B}^T]$, whose minimization with respect to \mathbf{B} yields an estimate of \mathbf{A}^{-1} .

³This property can also be derived from the Hadamard inequality which states that $\det \mathbf{R} \leq \det \text{diag}\mathbf{R}$ with equality if and only if \mathbf{R} is diagonal, see for ex. [13].

In practice, however, the covariance matrices $\mathbf{R}(t)$ of $\mathbf{X}(t)$ are unknown. Therefore a sensible approach is to replace them by some non parametric estimator. We consider a kernel estimator for $\mathbf{R}(t)$ (see for ex. [16], p. 25) *i.e.*

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

where $k(\cdot)$ is a positive kernel function and M is a parameter controlling the window width. The denominator $\sum_{s=1}^T k(\frac{t-s}{M})$ ensures that above right hand side is a weighted average of $\mathbf{X}(s) \mathbf{X}(s)^T$, but this factor is unimportant in our problem because it will cancel out. The separation procedure then consists in minimizing $\frac{1}{T} \sum_{t=1}^T \text{off}[\mathbf{B} \hat{\mathbf{R}}(t) \mathbf{B}^T]$ with respect to \mathbf{B} . But as $\hat{\mathbf{R}}(t)$ should vary slowly with t , because of the smoothing effect inherent in its definition and because of the slow variation of $\mathbf{R}(t)$, one may approximate the above criterion by

$$C_{MI} = \frac{1}{L} \sum_{l=1}^L \text{off}[\mathbf{B} \hat{\mathbf{R}}(lT/L) \mathbf{B}^T] \quad (15)$$

with L being some integer not exceeding T (the definition of $\hat{\mathbf{R}}(t)$ allows for non integer t .)

The minimization of (15) again amounts to the joint approximate diagonalization of a set of L matrices. In practice L can be chosen much smaller than T , to reduce cost. There is little to gain by taking large value for L , since then the successive matrices $\hat{\mathbf{R}}(lT/L)$ would be very similar.

D. Discussion

D.1 Connection between likelihood and mutual information

One can see that the Gaussian mutual information approach leads to a separation procedure which covers the one resulting from the block Gaussian likelihood approach as a special case, if the subintervals T_l have equal length. It is not hard to modify the former approach by allowing for varying window width, so that the case where the subintervals T_l are not of equal length are covered too, but we are not interested in such a level of generality.

It is not a coincidence however, that the two approaches lead to similar separating procedures. To elaborate on this point, we argue that for large T , the likelihood criterion C_{ML} should approach its expectation which, using (3), can be written as

$$EC_{ML} = \frac{1}{T} \sum_{t=1}^T D\{\mathbf{R}(t) | \mathbf{A} \boldsymbol{\Sigma}^2(t) \mathbf{A}^T\} + \text{Constant} \quad (16)$$

where $\mathbf{R}(t) = E[\mathbf{X}(t) \mathbf{X}(t)^T]$ is the true covariance matrix of $\mathbf{X}(t)$. We shall now use the notation \mathbf{B} for \mathbf{A}^{-1} to emphasize that it is a generic value of the parameter (not to be confused with the inverse of the true mixing matrix. Then using $D\{\mathbf{R} | \mathbf{A} \boldsymbol{\Sigma}^2 \mathbf{A}^T\} = D\{\mathbf{B} \mathbf{R} \mathbf{B}^T | \boldsymbol{\Sigma}^2\}$ and equality (12), the sum in (16) can be decomposed, reasoning as in section II-B, into

$$\frac{1}{T} \sum_{t=1}^T D\{\mathbf{B} \mathbf{R}(t) \mathbf{B}^T | \text{diag}[\mathbf{B} \mathbf{R}(t) \mathbf{B}^T]\} + \frac{1}{T} \sum_{t=1}^T D\{\text{diag}[\mathbf{B} \mathbf{R}(t) \mathbf{B}^T] | \boldsymbol{\Sigma}^2(t)\}.$$

Thus the likelihood measures (asymptotically) both terms in the above right hand side while the mutual information approach tries to minimize only the first term. However, in the first approach, if the values of $\sigma_k^2(t)$ were allowed to vary freely, the second term would reduce to zero, leaving only the first term. But one cannot really treat the $\sigma_k^2(t)$ as free parameters since one actually maximizes C_{ML} and not EC_{ML} and, without regularization, the solution would degenerate if there are too many parameters. The Gaussian

mutual information approach avoids this difficulty by taking the expectation first and then replacing the unknown covariance matrix $\mathbf{R}(t)$ by a non parametric estimate before performing the minimization.

We can also compare the solutions on the basis of the corresponding estimating equations. Denoting by $\hat{S}_i(\tau)$ the i -th component of $\hat{\mathbf{B}}\mathbf{X}(\tau)$, the minima of C_{MI} are easily shown to be solution of

$$\frac{1}{L} \sum_{l=1}^L \widehat{S}_i \widehat{S}_j \left(l \frac{T}{L} \right) / \widehat{S}_i \widehat{S}_i \left(l \frac{T}{L} \right) = 0, \quad 1 \leq i \neq j \leq K \quad (17)$$

where

$$\widehat{S}_i \widehat{S}_j(t) = \sum_{\tau=1}^T k \left(\frac{t-\tau}{M} \right) \hat{S}_i(\tau) \hat{S}_j(\tau) / \left[\sum_{\tau=1}^T k \left(\frac{t-\tau}{M} \right) \right].$$

The above equations are quite similar to (6) except that $S_i(t)S_j(t)$ and $\sigma_i^2(t)$ are replaced by local averages of $S_i S_j$ and of S_i^2 around the time point t and that the time average in (17) is sparser, using a time step of T/L instead of 1. The replacement of $S_i(t)S_j(t)$ by a local average should have no appreciable consequence, since in any case it is followed by a global average. The same can be said about the sparser sum because the local average of $S_i S_j$, as a function of time, is —by construction— slowly varying.

A more general procedure could be to solve (6) with $\sigma_i^2(t)$ estimated by

$$\hat{\sigma}_i^2(t) = \frac{\sum_{\tau=1}^T k \left(\frac{t-\tau}{M} \right) \hat{S}_i^2(t)}{\sum_{\tau=1}^T k \left(\frac{t-\tau}{M} \right)}. \quad (18)$$

This is more flexible than minimizing C_{MI} or C_{ML} , since the window parameter M in the estimator (18) can be made data driven, that is for each i , there could be a different parameter M_i which is adapted to the reconstructed sources sequence $\hat{S}_i(t)$. However, the later approaches, being based on the minimization of a criterion, offer two advantages over the former, which is based on a system of estimating equations. Firstly, such a system often has several solutions and hence there is a risk of finding a “spurious solution”. Secondly, a minimization algorithm can be controlled to ensure that the criterion decreases at each iteration and thus would at least converge to a local minimum. By contrast an iterative method for finding the solution of a system of estimating equations (such as the Newton method) may fail to converge if it is initialized too far away from the solution. The choice of the window parameter M is not crucial anyway, since one can tolerate some bias in the estimation of $\sigma_i^2(t)$ because — as mentioned in the introduction— the procedure still works even if the estimate is non consistent.

D.2 Super efficiency

In the noise free non stationary setting there is room for ‘super efficiency’, that is, for estimating the mixing matrix with an error which decreases faster than $1/\sqrt{T}$ as $T \rightarrow \infty$. Assume that the i -th source is silent over a given interval \mathcal{T} , while the other sources are not always silent over this interval:

$$\forall t \in \mathcal{T} S_i(t) = 0 \quad \text{and} \quad \forall j \neq i \exists t \in \mathcal{T} S_j(t) \neq 0. \quad (19)$$

Then it exists a vector \mathbf{b}_i such that $\mathbf{b}_i^T \mathbf{X}(t) = 0$ for all t in this interval. Since this vector must be orthogonal to all the columns of \mathbf{A} but the i -th column, it is proportional to the i -th row of \mathbf{A}^{-1} . Thus, in the situation described by (19), the i th row of \mathbf{A}^{-1} can be estimated *without* error from a finite number of samples.

The possibility of an error free estimation is preserved when the data in interval \mathcal{T} are summarized by the sample covariance matrix $\hat{\mathbf{R}}_{\mathcal{T}} = (\#\mathcal{T})^{-1} \sum_{t \in \mathcal{T}} \mathbf{X}(t) \mathbf{X}(t)^T$. This is because vector \mathbf{b}_i also is the unique (up to scale) solution of the equation $\mathbf{b}_i^T \hat{\mathbf{R}}_{\mathcal{T}} \mathbf{b}_i = 0$. Note that matrix $\hat{\mathbf{R}}_{\mathcal{T}}$, although subject to

estimation errors always have 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 where (19) holds is unlikely to occur (neither can we expect true noise-free instantaneous mixtures). But it is a guarantee of statistical effectiveness of an algorithm that it is capable of super efficiency when such a possibility exists. In particular, minimizing the criterion C_{ML}^* yields super-efficient estimates (even though not immediately obvious from expression (14)) whenever it happens that a given source is silent over one of the subintervals.

III. ALGORITHMS

A. Joint approximate diagonalization of positive matrices

The block Gaussian ML technique described at section II-B amounts to minimizing the criterion (14). It can be efficiently implemented thanks to an algorithm for the joint approximate diagonalization of several positive matrices which computes a matrix \mathbf{B} minimizing $C(\mathbf{B}) = \sum_{l=1}^L w_l \text{off}(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^T)$ where $\hat{\mathbf{R}}_1, \dots, \hat{\mathbf{R}}_L$ are positive matrices and w_1, \dots, w_L are positive weights. We give here only a brief description of the algorithm (for a full description and a proof of convergence, see the technical report [17] and a forthcoming paper [18]). Note that Flury and Gautschi [19], in a different context, has considered the same problem of minimizing $C(\mathbf{B})$, but the matrix \mathbf{B} is constrained to be orthogonal.

Our algorithm uses the classic Jacobi approach of operating by 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 and \mathbf{B}_j be a pair of rows of \mathbf{B} , the algorithm changes \mathbf{B} into a new matrix with these rows given by

$$\begin{bmatrix} \mathbf{B}_i \\ \mathbf{B}_j \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{B}_i \\ \mathbf{B}_j \end{bmatrix} - \mathbf{T}_{ij} \begin{bmatrix} \mathbf{B}_i \\ \mathbf{B}_j \end{bmatrix}, \quad (20)$$

the other rows being unchanged. Here, \mathbf{T}_{ij} is a 2×2 matrix which can be chosen such that the criterion is sufficiently decreased (in a sense to be specified). 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.

A key point is that the transformation matrix \mathbf{T}_{ij} in (20) can be computed in closed form as follows. Define the quantities:

$$g_{ij} = \sum_{l=1}^L w_l \frac{(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^T)_{ij}}{(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^T)_{ii}}, \quad \omega_{ij} = \sum_{l=1}^L w_l \frac{(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^T)_{jj}}{(\mathbf{B}\hat{\mathbf{R}}_l\mathbf{B}^T)_{ii}}, \quad \begin{bmatrix} h_{ij} \\ h_{ji} \end{bmatrix} = \begin{bmatrix} \omega_{ij} & 1 \\ 1 & \omega_{ji} \end{bmatrix}^{-1} \begin{bmatrix} g_{ij} \\ g_{ji} \end{bmatrix} \quad (21)$$

where we have assumed that $\sum_{l=1}^L w_l = 1$ (if this is not the case, the weights should be thus normalized) and build \mathbf{T}_{ij} as

$$\mathbf{T}_{ij} = \frac{2}{1 + \sqrt{1 - 4h_{ij}h_{ji}}} \begin{bmatrix} 0 & h_{ij} \\ h_{ji} & 0 \end{bmatrix}. \quad (22)$$

The transformation (20) with \mathbf{T}_{ij} given by (22) always decreases the criterion $C(\mathbf{B})$ unless $g_{ij} = g_{ji} = 0$. This is very significant since g_{ij} is, for $i \neq j$, nothing but the (i, j) -th element of the relative gradient matrix of $C(\mathbf{B})$. For a proof and a convergence analysis of the algorithm, see [17].

Note that the algorithm does not try to control explicitly the scales of the recovered sources (the criterion $C(\mathbf{B})$ itself is scale invariant) but rescaling could be done after convergence if desired. Numerical problems due to large imbalance in scales is not a concern thanks to the very fast convergence (see experimental section) of the procedure.

The use of the above algorithm for the minimization for the criterion (14) is referred to as ‘JD-BGL’ (joint diagonalization for the block Gaussian likelihood). Of course, this algorithm can also be used to minimize (15), but for simplicity this more general criterion will not be considered in our simulations.

B. On-line algorithms

B.1 Simple stochastic gradient

The simplest idea for on-line separation is to implement a stochastic relative gradient algorithm for the minimization of the likelihood criterion. The relative gradient of this criterion likelihood is given by (5). Reasoning as in [11], this suggests the following algorithm for updating a separating matrix $\mathbf{B}(t)$ upon reception of a new sample $\mathbf{X}(t)$:

$$\hat{\mathbf{B}}(t+1) = \hat{\mathbf{B}}(t) - \lambda \mathbf{G}(t) \hat{\mathbf{B}}(t) \quad (23)$$

where λ is a small positive constant and $\mathbf{G}(t)$ is the relative stochastic gradient:

$$\mathbf{G}(t) = \hat{\Sigma}^{-2}(t) \hat{\mathbf{S}}(t) \hat{\mathbf{S}}(t)^T - \mathbf{I} \quad \text{with} \quad \hat{\mathbf{S}}(t) = \mathbf{B}(t) \mathbf{X}(t). \quad (24)$$

Here $\hat{\Sigma}^2(t) = \text{diag}[\hat{\sigma}_1^2(t), \dots, \hat{\sigma}_K^2(t)]$ where each $\hat{\sigma}_k^2(t)$ is some non parametric estimate of $\sigma_k^2(t)$, for instance the output of an exponential filter applied to the sequence $\hat{S}_k^2(t)$:

$$\hat{\sigma}_k^2(t) = \hat{\sigma}_k^2(t-1) + \rho [\hat{S}_k^2(t) - \hat{\sigma}_k^2(t-1)]. \quad (25)$$

where ρ is a small positive learning step. Other filters could be used provided they produce positive output for any positive input. But the most important point is that λ must be significantly smaller than ρ since the estimated separating matrix $\hat{\mathbf{B}}$ should be nearly constant in a large range of time in which the source variances and hence their estimates can vary significantly.

This type of algorithm closely parallels those derived in the non Gaussian stationary case where the quantity $\hat{\Sigma}^{-2}(t) \hat{\mathbf{S}}(t)$ in (24) is replaced by a function $\phi[\hat{\mathbf{S}}(t)]$ which still operates component-wise on $\hat{\mathbf{S}}(t)$ but is (i) non linear (non Gaussianity) and (ii) independent of t (stationarity). These simple relative gradient algorithms are well behaved but they can be significantly improved (in terms of convergence speed) at little additional cost by a Newton-like procedure.

B.2 A Newton-like on-line technique

For a given separating matrix \mathbf{B} and for λ a small positive parameter, consider an exponentially weighted relative gradient matrix $\bar{\mathbf{G}}_t(\mathbf{B})$ similar to (5):

$$\bar{\mathbf{G}}_t(\mathbf{B}) = \sum_{\tau \leq t} \lambda (1 - \lambda)^{t-\tau} \Sigma^{-2}(\tau) \mathbf{B} \mathbf{X}(\tau) \mathbf{X}(\tau)^T \mathbf{B}^T - \mathbf{I} \quad (26)$$

computed at time t based on the past samples. As before, $\Sigma^2(\tau)$ denotes the diagonal matrix with diagonal elements $\sigma_1^2(\tau), \dots, \sigma_K^2(\tau)$ and we start by assuming that the source variances are known; they are later replaced by some estimates. Our plan is to solve $\bar{\mathbf{G}}_t[\hat{\mathbf{B}}(t)] = 0$ assuming that the relative gradient (26) already vanishes at time $t-1$, that is, $\bar{\mathbf{G}}_{t-1}[\hat{\mathbf{B}}(t-1)] = 0$. Since λ is small, the solution $\hat{\mathbf{B}}(t)$ at time t would differ from $\hat{\mathbf{B}}(t-1)$ by a term of order λ , which —similarly to (23)— we write as a relative variation:

$$\hat{\mathbf{B}}(t) = \hat{\mathbf{B}}(t-1) - \lambda \mathbf{H}(t) \hat{\mathbf{B}}(t-1). \quad (27)$$

Matrix $\mathbf{H}(t)$ can be obtained (at first order in λ) by computing the first order expansion of the gradient. One finds that if $\tilde{\mathbf{G}}_{t-1}(\mathbf{B}) = 0$, then (see appendix for details)

$$\begin{aligned} \tilde{\mathbf{G}}_t(\mathbf{B} - \lambda\mathbf{H}\mathbf{B}) &= \lambda[\boldsymbol{\Sigma}^{-2}(t)\mathbf{B}\mathbf{X}(t)\mathbf{X}(t)^\top\mathbf{B}^\top - \mathbf{I}] - \lambda\mathbf{H}^\top \\ &\quad - \lambda\sum_{\tau \leq t} \lambda(1-\lambda)^{t-\tau}\boldsymbol{\Sigma}^{-2}(\tau)\mathbf{H}\mathbf{B}\mathbf{X}(\tau)\mathbf{X}(\tau)^\top\mathbf{B}^\top + O(\lambda^2). \end{aligned} \quad (28)$$

A reasonable approximation to the last term in (28) is obtained when \mathbf{B} is close to the true separating matrix so that, owing to the smoothing effect of the exponential average, the factor $\mathbf{B}\mathbf{X}(\tau)\mathbf{X}(\tau)^\top\mathbf{B}^\top$ can be replaced by $\boldsymbol{\Sigma}^2(\tau)$. Thus, the sum on the right hand side of (28) can be approximated by $-\lambda\sum_{\tau \leq t} \lambda(1-\lambda)^{t-\tau}\boldsymbol{\Sigma}^{-2}(\tau)\mathbf{H}(t)\boldsymbol{\Sigma}^2(\tau)$. Neglecting the $O(\lambda^2)$ term in (28), the condition $\tilde{\mathbf{G}}_t(\mathbf{B} - \lambda\mathbf{H}\mathbf{B}) = 0$ becomes a linear equation in \mathbf{H} :

$$\mathbf{H}^\top + \sum_{\tau \leq t} \lambda(1-\lambda)^{t-\tau}\boldsymbol{\Sigma}^{-2}(\tau)\mathbf{H}\boldsymbol{\Sigma}^2(\tau) = \boldsymbol{\Sigma}^{-2}(t)\mathbf{B}\mathbf{X}(t)\mathbf{X}(t)^\top\mathbf{B}^\top - \mathbf{I}. \quad (29)$$

Substituting \mathbf{B} by $\mathbf{B}(t-1)$ and putting $\hat{S}_i(t) = [\mathbf{B}(t-1)\mathbf{X}(t)]_i$, the matrix equation (29) can be decomposed as:

$$h_{ji} + h_{ij} \sum_{\tau \leq t} \lambda(1-\lambda)^{t-\tau} \sigma_j^2(\tau) / \sigma_i^2(\tau) = \hat{S}_i(t)\hat{S}_j(t) / \sigma_i^2(t), \quad 1 \leq i \neq j \leq K \quad (30)$$

$$2h_{ii} = \hat{S}_i^2(t) / \sigma_i^2(t) - 1, \quad 1 \leq i \leq K, \quad (31)$$

where h_{ij} denotes the (i, j) entry of \mathbf{H} . The equations (31) control the scales of the recovered sources. However, such a control is not required because of an invariance property discussed below.

In practice, the source variance $\sigma_k^2(t)$ is obtained by an on-line estimator $\hat{\sigma}_k^2(t)$ with a learning step ρ , like the one defined by (25) and the sum in (30) is obtained by an exponential smoothing of $\hat{\sigma}_j^2(t) / \hat{\sigma}_i^2(t)$ with a learning step λ . This yields the following on-line algorithm.

1. Compute $\hat{\mathbf{S}}(t) = \mathbf{B}(t-1)\mathbf{X}(t)$, update $\hat{\sigma}_k^2(t)$ by (25) and update $\hat{\omega}_{ij}(t)$ 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 (27) where the diagonal of matrix $\mathbf{H}(t)$ is set to zero and its off diagonal elements are the solutions of (30) *i.e.*

$$\begin{bmatrix} h_{ij}(t) \\ h_{ji}(t) \end{bmatrix} = \begin{bmatrix} \hat{\omega}_{ij}(t) & 1 \\ 1 & \hat{\omega}_{ji}(t) \end{bmatrix}^{-1} \begin{bmatrix} \hat{S}_i(t)\hat{S}_j(t) / \hat{\sigma}_i^2(t) \\ \hat{S}_j(t)\hat{S}_i(t) / \hat{\sigma}_j^2(t) \end{bmatrix} \quad (32)$$

As before, the parameter λ should be much smaller than ρ . The 2×1 vector on the right hand side of (32) contains the entries (ij) and (ji) of the relative stochastic gradient $\mathbf{G}(t)$ of (24) so that the 2×2 matrix in (32) plays the role of the Hessian in a Newton algorithm. In this respect, we must stress that this matrix is guaranteed to remain definite positive if the sequences $\hat{\sigma}_i^2(t)$ and $\hat{\sigma}_j^2(t)$ are not proportional (this is because $\hat{\omega}_{ij}(t)\hat{\omega}_{ji}(t) \geq 1$ which is proved by applying the Cauchy-Schwartz inequality to the sequences $\{u_{ij}(\tau)\}_{\tau \leq t}$ and $\{u_{ji}(\tau)\}_{\tau \leq t}$ where $u_{ij}(\tau) = \sqrt{\lambda(1-\lambda)^{t-\tau}} \hat{\sigma}_j(\tau) / \hat{\sigma}_i(\tau)$).

We have simply set $h_{ii}(t) = 0$ because the effect of the diagonal terms of $\mathbf{H}(t)$ is to change the scale of $\hat{S}_i(t)$. By taking $h_{ii}(t) = 0$, we choose not to update $\hat{\mathbf{B}}$ in the directions which change the scales of each output $\hat{S}_i(t)$ hence not to try to control the scale of the outputs. This is not in contradiction with the approximation that the covariance matrix of $\hat{\mathbf{S}}(t)$ is close to $\boldsymbol{\Sigma}^2(t)$ because the variance profile $\sigma_i^2(t)$ is actually estimated from $\hat{S}_i(t)$ itself.

Further, we note that the algorithm is scale-invariant in the following sense: for a given data sequence $\{\mathbf{X}(t)\}$, if $\{\mathbf{B}(t)\}$ is a trajectory of the algorithm, another possible trajectory is $\{\Lambda\mathbf{B}(t)\}$ where Λ is any invertible diagonal matrix. This is because in such a case the values of $\hat{\mathbf{S}}(t)$ are rescaled into $\Lambda\hat{\mathbf{S}}(t)$ leading to a rescaling of \mathbf{H} into $\Lambda\mathbf{H}\Lambda^{-1}$. This invariance is an *exact* property but the lack of control of the scales may become a problem in applications (due to numerical errors) if the system is always in learning mode, in which case there is a real danger of a slow, continuous drift of the scales. In such a case, one may wish to control the scales by setting $\mathbf{H}_{ii}(t) = \alpha[\hat{S}_i^2(t) - 1]$ for $i = 1, \dots, K$ in order to drive gently to 1 the long term average variance of each estimated source signal. Note that equation (31) suggests to take $\alpha = 0.5$: by choosing smaller values of α , the ‘scale learning speed’ is made proportionally slower.

B.3 On-line versions of the joint diagonalization algorithm

The block Gaussian approach can be easily turned into a ‘block on-line’ algorithm. In this context, instead of having a fixed number of data samples, one has a stream of them, but it can be subdivided into data blocks of a given length, say m . For the l -th data block, one can compute the sample covariance matrix $\hat{\mathbf{R}}_l$ similarly to (8). The L most recent covariance matrices are kept in memory and, after block l has become available, one could perform the joint approximate diagonalization of the matrices $\hat{\mathbf{R}}_l, \dots, \hat{\mathbf{R}}_{l+1-L}$ 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. In this case, one should store the covariance matrices of the estimated source signals $\hat{\mathbf{S}}(t)$.

This block on-line version of JD-BGL is implemented as follows. At a given point in time, the $(l-1)$ -th block of samples has been received, the current value of the separating matrix is $\mathbf{B}^{(l-1)}$ and L covariance matrices $\hat{\mathbf{R}}^{(l-1)}, \dots, \hat{\mathbf{R}}^{(l-L)}$ are stored in memory. Upon reception of the next m samples (those forming the l -th block),

1. Drop the oldest matrix, that is $\hat{\mathbf{R}}^{(l-L)}$, and store the sample covariance matrix $\hat{\mathbf{R}}^{(l)}$ of the current source estimates $\mathbf{B}^{(l-1)}\mathbf{X}(t)$ for $\mathbf{X}(t)$ belonging to the l -th block.
2. Compute a $K \times K$ transformation $\mathbf{T}^{(l)}$ by one sweep of the joint diagonalization algorithm applied to the L matrices $\hat{\mathbf{R}}^{(l)}, \dots, \hat{\mathbf{R}}^{(l+1-L)}$. During this sweep, each matrix $\mathbf{R}^{(n)}$ is updated into $\mathbf{T}^{(l)}\mathbf{R}^{(n)}\mathbf{T}^{(l)\text{T}}$ for $n = l+1-L, \dots, l$ and the separating matrix is updated into $\mathbf{B}^{(l)} = \mathbf{T}^{(l)}\mathbf{B}^{(l-1)}$.

Likewise, the Gaussian mutual information approach 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 and therefore we shall use the notation $\hat{\mathbf{R}}(t)$ instead. In the on-line processing context, this is best done by applying a low-pass filter to the matrix sequence $\mathbf{X}(t)\mathbf{X}(t)^\text{T}$. One should however be careful to ensure that the output of the filter be positive matrices. A simple low pass filter which meets this requirement is the exponential filter, defined by

$$\hat{\mathbf{R}}(t) = \hat{\mathbf{R}}(t-1) + \rho[\mathbf{X}(t)\mathbf{X}(t)^\text{T} - \hat{\mathbf{R}}(t-1)] \quad (33)$$

where ρ is a positive number less than 1. The separating matrix \mathbf{B} is then obtained (at time t) by jointly approximately diagonalizing the matrices $\hat{\mathbf{R}}(t), \hat{\mathbf{R}}(t-m), \dots, \hat{\mathbf{R}}(t+m-mL)$. Here the role of the integer m is to reduce the number of matrices to be diagonalized since the product $(m-1)L$ must be large enough so that the source variance can vary significantly in an interval of length $(m-1)L$. Further, m must not be too large so that $\hat{\mathbf{R}}(t)$ does not change much in an interval of length m . Typically, when $\hat{\mathbf{R}}(t)$ is obtained by (33), m should be inversely proportional to ρ (for example $m = 1/(2\rho)$) because $1/(2\rho)$ is the equivalent window width of the exponential low pass filter. Hence $\hat{\mathbf{R}}(t)$ cannot change significantly over periods of time shorter $1/(2\rho)$ but could change over longer periods.

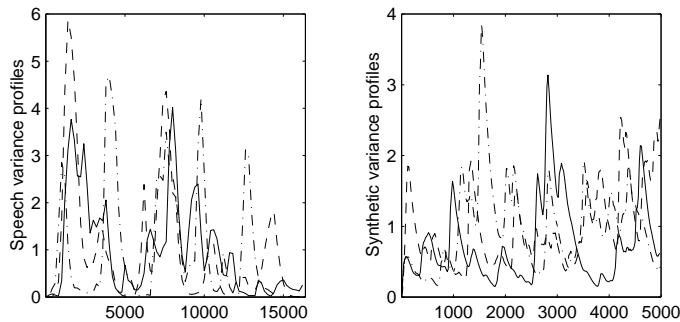


Fig. 1. Typical variance profiles of the signals used in the experiments. Left: (a) the three speech waveforms. Right: (b) the synthetic signals.

Again, it should be noted that in the joint approximate diagonalization algorithm, if one starts at the value of the separating matrix obtained at the previous step, then convergence is reached very quickly since the matrices to be diagonalized have not changed much from one time step to the next. Typically, one sweep of the algorithm is expected to be enough. The resulting on-line algorithm is:

1. Compute $\hat{\mathbf{R}}(t)$ according to (33)
2. Apply one sweep of the joint approximate diagonalization algorithm to the matrices $\hat{\mathbf{R}}(t)$, $\hat{\mathbf{R}}(t-m)$, \dots , $\hat{\mathbf{R}}(t+m-mL)$, starting with the previous estimate $\hat{\mathbf{B}}(t-1)$, to obtain the new estimate $\hat{\mathbf{B}}(t)$ of the inverse of the mixing matrix.

IV. NUMERICAL EXPERIMENTS

Signals. Our experiments use both synthetic and real source signals. The real signals are three speech waveforms sampled at 8 kHz. About two seconds of speech are available for each speaker (one male, one female, one child).⁴ The variance profiles, estimated as in (25) with $\rho = 5 \cdot 10^{-3}$, are displayed on figure 1.a. Regarding the synthetic signals, they are drawn according to our working assumptions: for each i , a smooth scale profile $\sigma_i(t)$ is first drawn and used to modulate an i.i.d. sequence of zero-mean unit-variance normal variables (see fig. 1.b for a sample of the corresponding estimated variance profiles.)

Equivariance. All the algorithms described in this paper are equivariant. This means that the behavior of the algorithm, in particular the accuracy of separation, is independent of the mixing matrix (see e.g. [1]). More precisely, for each algorithm, the distribution of the global system $\hat{\mathbf{B}}\mathbf{A}$ (or the trajectory of $\hat{\mathbf{B}}(t)\mathbf{A}$ for on-line methods) does not depend on \mathbf{A} but only on the distribution of the sources.

A. Batch algorithms

Local minima. A simple test for the existence of ill minima has been conducted as follows. We generate random 3×3 mixing matrices \mathbf{A} by drawing independently each of their coefficients under a zero mean unit variance normal distribution and apply them to the three speech signals. A separating matrix \mathbf{B} is computed by applying the JD-BGML algorithm with window length $\#T_l = 100$. The experience has been repeated 1000 times. The very same global system $\mathbf{B}\mathbf{A}$ has been found each time.

Speed of convergence. Only a few sweeps are necessary for the convergence of JD-BGL. Since the algorithm converges quite quickly, there is no need for a sophisticated stopping rule. In our experiments, we stop when the gradient has reached a numerically small —as opposed to ‘statistically small’— value. This incurs little extra sweeps because convergence is very fast in the final phase.

⁴Thanks to O. Cappé for providing us with these signals

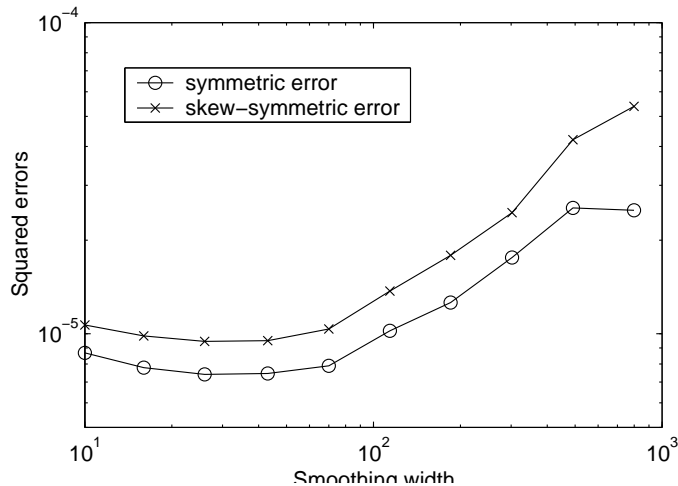


Fig. 2. Accuracy versus window length

Stationarity length. In all the algorithms, there is a parameter (block length, learning step, etc) which should be adapted to the ‘stationarity length’ of the source signals. For the speech signals, we can use prior knowledge about the average duration of phonemes and have the block length (or $1/\rho$ in the on-line algorithms) scaling as a fraction of the stationarity length. In figure 2, we investigate how the estimation accuracy depends on the block length. We use synthetic signals with a scale profiles as shown on fig. 1 where the stationarity length is about 300 samples. Define a relative error for the (i, j) pair as $e_{ij} = (\hat{\mathbf{A}}^{-1} \mathbf{A})_{ij}$, $i \neq j$, and the symmetric (resp. skew symmetric) squared error as $e_{\pm}(\hat{\mathbf{A}}) = \frac{1}{2} \sum_{i \neq j} (e_{ij} \pm e_{ji})^2$. We evaluate the mean squared errors $\text{MSE}^{\pm} = \frac{1}{N} \sum_{k=1}^N e_{\pm}(\hat{\mathbf{A}})$ over $N = 1000$ runs for $T = 8000$ samples. The scale profiles are fixed through all the runs and they modulate an i.i.d. Gaussian process which is different for each run. The results are displayed on figure 2 for the JD-BGL algorithm as a function of the block size. It is seen that the skew-symmetric error MSE^{-} dominates the overall error and that both types of errors remain close to their lowest level as long as the block length remains shorter than the stationarity length. Underestimating the stationarity length, that is, selecting a much shorter block length, does not seem to hurt very much the accuracy. However, it increases the computational cost by increasing proportionally the number of covariance matrices to be diagonalized.

Super efficiency. The separating matrix obtained on the speech signals with the JD-BGL algorithm is such that

$$\hat{\mathbf{B}}\mathbf{A} = \begin{bmatrix} 1.00000 & -0.00024 & 0.00007 \\ 0.00008 & 1.00000 & 0.00003 \\ 0.00014 & -0.00026 & 1.00000 \end{bmatrix} \quad (34)$$

after fixing the indeterminations to obtain a unit diagonal. The very small off-diagonal terms are to be explained by a ‘super efficiency effect’ as discussed at section II-D and by the fact that speech often includes some short periods of silence.

In order to further illustrate the super efficiency effect, we have set up an experiment with synthetic signals of the form $S_i(t) = a_i(t)n_i(t)$ (with $K = 3$ sources and $T = 1000$ samples) where the sequences $n_i(t)$ are drawn Gaussian i.i.d. and $a_i(t)$ is a deterministic envelope in the form $a_i(t) = \sigma + b_i(t)$. For $i = 1, 2$, we take $b_i(t)$ to have many periods of silence while the values of $b_3(t)$ do not reach 0. By varying σ , we can easily adjust the minimum value of the local variances. In particular, we can make sources 1 and 2 be ‘silent at level σ ’ with the ‘amplitude floor’ σ as small as desired. In figure 3, we plot $|(\hat{\mathbf{B}}\mathbf{A})_{ij}|$

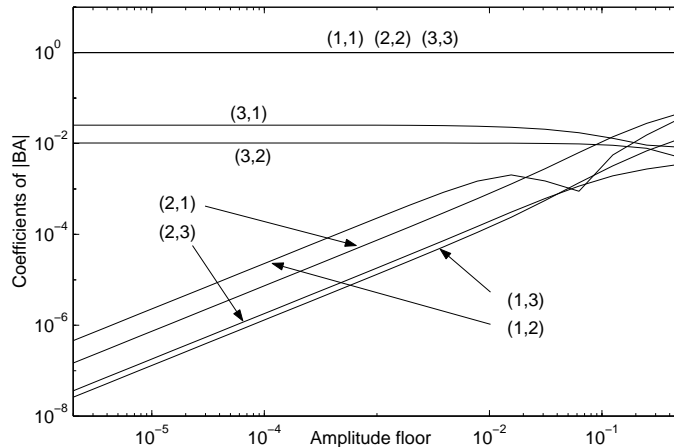


Fig. 3. Super efficient separation by the JD-BGL technique of sources with periods of silence. The plot shows $|(\hat{\mathbf{B}}\mathbf{A})_{ij}|$ for $1 \leq i, j \leq K$ for $K = 3$ sources as a function of an amplitude floor σ . For two sources out of three ($i = 1, 2$), the variance profiles have periods of silence at level σ (see text). The nine curves (three of them overlap) are labelled with the index pair (i, j) . The estimation error $|(\hat{\mathbf{B}}\mathbf{A})_{ij}|$ goes to 0 when $\sigma \rightarrow 0$ for $i = 1, 2$: the first and second sources are recovered super-efficiently in this limit. See text for details.

versus the amplitude floor σ with $\hat{\mathbf{B}}$ estimated by the JD-BGL algorithm. For the sake of comparison, the usual indeterminations of permutation and scale are fixed in such a way that $(\hat{\mathbf{B}}\mathbf{A})_{ii} = 1$ so that three curves ($i = 1, 2, 3$) overlap at level 1 in the figure. The figure shows that when $\sigma \rightarrow 0$, the values of $(\hat{\mathbf{B}}\mathbf{A})_{ij}$ converge to a non-zero value for $i = 3$ while they still decrease (the figure suggests a decrease proportional to σ) for $i = 1, 2$. This experiment confirms that ‘the more silent a source, the more accurately it can be separated’. As discussed above, perfect extraction is possible in theory with a finite number of samples (super efficiency effect) for sources which are truly silent ($\sigma = 0$) on an interval.

Other block algorithms. Since speech is definitely not normally distributed, it is also possible to consider traditional techniques derived from non Gaussian modeling. As an example, the global system $\hat{\mathbf{B}}\mathbf{A}$ obtained with the JADE algorithm [14] algorithm is

$$\hat{\mathbf{B}}\mathbf{A} = \begin{bmatrix} 1.000 & -0.003 & 0.005 \\ -0.039 & 1.000 & 0.007 \\ -0.008 & -0.001 & 1.000 \end{bmatrix} \quad (35)$$

which is significantly below the quality (34) of JD-BGL. This is not to be attributed to the use of fourth-order cumulants in JADE, as seen from the results of the following hybrid between JADE and JD-BGL: the observed signals are whitened with a matrix \mathbf{W} which is —by a standard argument— supposed to turn the mixing matrix into an orthogonal matrix $\mathbf{U} = \mathbf{W}\mathbf{A}$ and matrix \mathbf{U} is estimated as the orthonormal matrix which minimizes the joint diagonality criterion

$$C(\mathbf{U}) = \sum_l \sum_{i \neq j} w_l (\mathbf{U}^T \hat{\mathbf{R}}_l \mathbf{U})_{ij}^2 \quad (36)$$

where $\hat{\mathbf{R}}_l$ is the sample covariance matrix of the whitened signals of the l -th interval. This algorithm has been proposed by Parra and Spence [10]; it uses the same statistics as JD-BGL and has a similar objective: minimizing output decorrelation expressed via a joint diagonalization criterion. It differs from JD-BGL in two respects: it uses a ‘naive’ (as opposed to likelihood-based) measure of diagonality and it

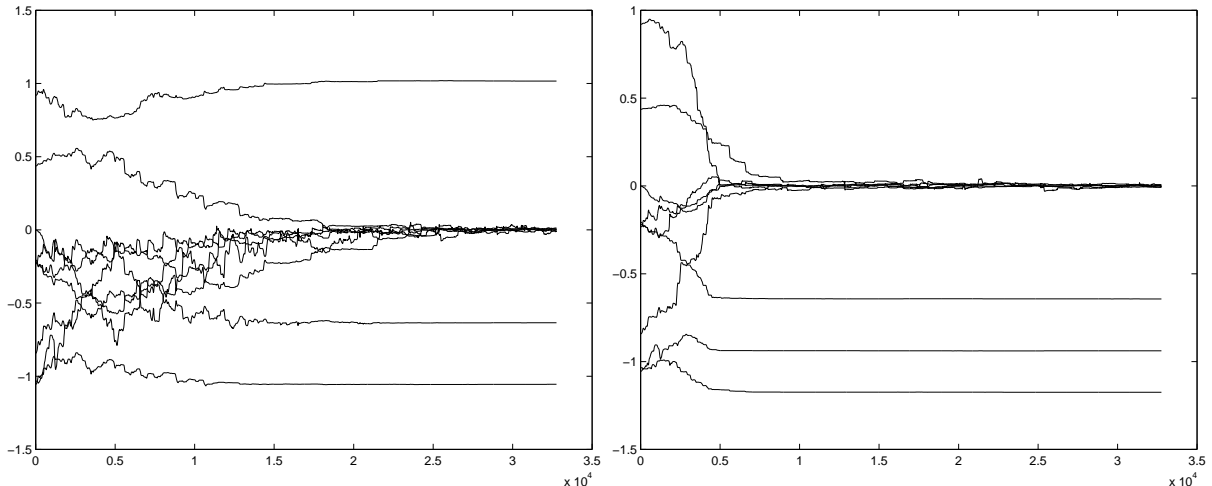


Fig. 4. Convergence of the 9 coefficients of the global system $\mathbf{B}(t)\mathbf{A}$ for a $K = 3$ source case. Left: the ‘regular’ relative gradient technique. Right: the Newton-like technique.

enforces the (empirical) decorrelation (or whiteness) of the recovered signals. Minimizing the quadratic criterion (36) yields:

$$\hat{\mathbf{B}}\mathbf{A} = \begin{bmatrix} 1.000 & 0.009 & -0.005 \\ -0.051 & 1.000 & 0.006 \\ 0.002 & -0.000 & 1.000 \end{bmatrix} \quad (37)$$

which is again mediocre when compared to (34). This shows that neither JADE nor BG-JADE are able to exploit the potential of super efficiency. They are also penalized by the fact that our speech signals are not well decorrelated empirically. Actually, their sample covariance matrix (after renormalization to unit variance) is found to be:

$$\frac{1}{T} \sum_{t=1}^T \mathbf{s}(t)\mathbf{s}(t)^T = \begin{bmatrix} 1.000 & 0.042 & 0.003 \\ 0.042 & 1.000 & -0.006 \\ 0.003 & -0.006 & 1.000 \end{bmatrix} \quad (38)$$

so that no algorithm based on prior whitening can be expected to do very well on this data set. Indeed, any matrix \mathbf{B} enforcing the decorrelation of its first and second output cannot be very close to a separating matrix since the *sample* correlation between source 1 and source 2 is not small at all but close to 4% according to matrix (38).

B. On-line algorithms

Stochastic relative gradient algorithms. Figure 4 shows the convergence of the 9 coefficients $(\mathbf{B}(t)\mathbf{A})_{ij}$ of the global system in a $K = 3$ scenario for on-line algorithms. The left panel shows the convergence for the ‘regular’ relative gradient algorithm (23) with the diagonal of the relative gradient $\mathbf{G}(t)$ set to 0. We have used synthetic signals as above and set $\rho = 10^{-2}$ and $\lambda = \rho/20$. The right panel shows the convergence of the Newton-like algorithm (27) with the same signals, same parameters and same starting point. The significantly faster convergence is clearly visible. The two runs shows different limiting values for the scales (the non-zero limits differ in each plot) which is to be expected since the two algorithms have a different policy for scale control.

Block on-line algorithms. Figure 5 shows an example of the convergence of the on-line version of JD-BGL. We use the speech signals, a block length of $m = 320$ samples (40 ms) and a memory of $L = 12$ matrices to be jointly diagonalized. The figure shows the convergence of the 9 coefficients the global

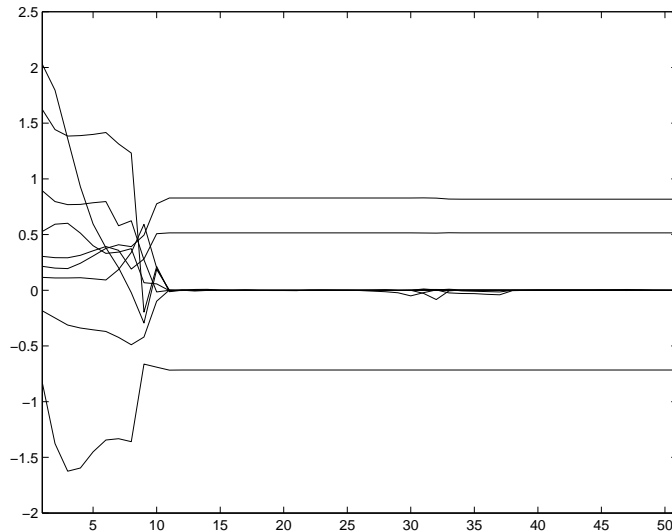


Fig. 5. Convergence of the on-line version of JD-BGL.

system \mathbf{BA} versus the number of blocks. In this figure, the convergence is reached after about 11 blocks, that is even before the memory is full. This is significantly faster than the on-line Newton-like algorithm.

V. DISCUSSION

Relations to previous works. This paper has some close connections to previous works. Most notably, Matsuoka *et al.* consider a criterion which is essentially the Gaussian mutual information (without relating it to the likelihood or to the mutual information itself) but they only propose a stochastic gradient technique for its optimization; this is bound to be much less efficient than our pseudo-Newton technique. More recently, Parra and Spence [10] proposed the criterion (36) for which an efficient minimization algorithm exists but which is not directly related to maximum likelihood. Another technique is considered by Ngo [9] based on the heuristic that a separating matrix can be found by requiring that its outputs are uncorrelated over successive intervals.

Souloumiac [6] and Tsatsanis [7], in very similar papers, consider the case when the interval $[1, T]$ is divided in only two subintervals. Then the joint diagonalization of $\hat{\mathbf{R}}_1$ and $\hat{\mathbf{R}}_2$ can be achieved exactly (since this is equivalent to solving a generalized eigenvalue problem) and a readily available algorithm can be used. As we have seen, this idea actually corresponds to ML estimation in a block Gaussian model with two sub-intervals.

We also want to point out the similarities between the ‘efficient approaches’ considered in this paper: the JD-BGL algorithm and the Newton-like algorithm of section III-B.2. In both cases, a key step is the transformation of the gradient g_{ij} into a ‘rectified gradient’ h_{ij} (see eqs. (21) and (32)). 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. Thus, it is likely that the ‘natural gradient’ approach of Amari [20] would result in similar algorithms. We note that on-line algorithm takes its particular simple form thanks to an approximation which is only valid when the model holds. When this is not the case, the algorithm is still well behaved because the gradient is rectified by a matrix which is always positive.

Non stationarity and non Gaussianity. It is interesting to compare the present Gaussian non stationary (GnS) model for the sources to the more traditional model of stationary non Gaussian (SnG) sources. In both cases, a rough estimate of the nuisance parameter (the distribution of the sources) is sufficient to construct a consistent estimate of the mixture. In both cases, this rough estimate determines a function to be applied to the estimated sources: the ‘score function’ which is the log-derivative of the probability density. In the SnG case, the score function is a fixed (time independent) non linear function; in the GnS the score function is linear time-varying (division by the local estimate of the variance). In the SnG case, it is often sufficient to have a broad idea about the distribution of the sources (for instance, that the source distributions have heavy tails). Therefore, when the type of sources to be extracted is known in advance, there might be no need to estimate the nonlinear functions from the data themselves. In contrast, in the GnS case, the local variances must be estimated from the data themselves because, in most realistic applications, it is not possible to determine in advance, even vaguely, what the variance profiles will be (speech processing being an obvious example).

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: The former ensures that decorrelation condition $E[S_i(t)S_j(t)/\sigma_i^2(t)] = 0$ (of which the estimating equation (6) is an empirical version) holds for zero-mean sources while the latter ensures that the local variances can be (roughly) estimated. However, a ‘slow variation of the variance profile’ is not –strictly speaking– related to the well defined notion of stationarity: assume, for instance, that the i -th source signal is $S_i(t) = a_i(t)n_i(t)$ where $n_i(t)$ is an i.i.d. sequence and $a_i(t)$ is a *stationary* process which is slowly varying in time with, say, a typical time constant τ . Such a model produces samples which are perfectly appropriate for our class of algorithms provided we can observe enough fluctuations of each $a_i(t)$, that is if T is (significantly) larger than τ . However, this model is, strictly speaking and by definition, a stationary model. Conversely, it is a simple matter to build non stationary processes having a constant variance. Our algorithms would fail to separate such sources since they only capture the non stationarity in amplitude.

In summary, it would be more accurate to describe our algorithms as applying to independent sources with ‘slow’ amplitude modulation.

A final comment regards a connection between non stationarity and non Gaussianity. For simplicity, consider again a simple non stationary model in which the i -th source sequence is $S_i(t) = a_i(t)n_i(t)$ with $n_i(t)$ an i.i.d. sequence of zero-mean unit variance normal variables and $a_i(t)$ a ‘slowly varying’ amplitude. 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 if the amplitude $a_i(t)$ varies significantly over $[1, T]$.

For the same reasons, Parra and Spence argues [21] that in ‘natural’ signal or images the non Gaussianity often is the result of ignoring non stationarity (in the same sense as in the above paragraph). He goes further by noting that a mixture of independent zero mean Gaussian variables with different variances always results in a density with positive kurtosis and takes this fact as an explanation of the fact that distributions with positive kurtosis (or more generally: heavy tailed distributions) are most often encountered in real signals.

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

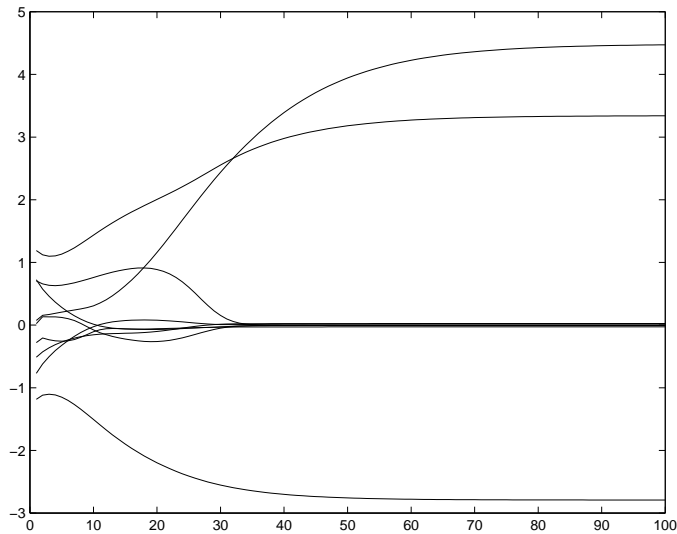


Fig. 6. Relative gradient algorithm based on the Cauchy score function (see text). Convergence of the 9 coefficients of the global system vs the number of iterations.

its posterior distribution given $\hat{S}_i(t)$. Such an estimate can easily be computed in closed form if a nice prior is used. With an inverse Gamma prior, this can also be seen [22] as assuming that n_0 extra data points are available with sample variance σ_0^2 , in which case the regularized variance estimate simply is $\hat{\sigma}_i^2(t) = (\hat{S}_i^2(t) + n_0\sigma_0^2)/(1 + n_0)$ and n_0 and σ_0^2 are free hyper parameters. The estimating equations become

$$\frac{1}{T} \sum_t \psi(\hat{S}_i(t)) \hat{S}_j(t) - \delta_{ij} = 0 \quad (39)$$

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 exactly 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. (39) amounts to using a model of i.i.d Cauchy sources. As an example, figure 6 shows the convergence of $(\mathbf{BA})_{ij}$ for the speech signals when equations (39) are solved by a relative gradient technique which updates an estimate of \mathbf{B} into $\mathbf{B} - \mu\mathbf{GB}$ where \mathbf{G}_{ij} is given by the left hand side of (39) and $\mu = 0.2$. After convergence, the global system is

$$\hat{\mathbf{B}}\mathbf{A} = \begin{bmatrix} 1.0000 & -0.0088 & -0.0058 \\ -0.0062 & 1.0000 & -0.0014 \\ -0.0019 & -0.0012 & 1.0000 \end{bmatrix}. \quad (40)$$

Again, the results are good but not as good as for JD-BGL: the simple estimating equations (39) fails to be ‘super efficient’. This is no surprise since the underlying model does not take into account the almost-singularity of the source distributions (namely the fact that there are periods of quasi-silence).

Conclusions. Non stationarity provides us with necessary information for the purpose of source separation, just as non Gaussianity does. An advantage of the second order methods exploiting non stationarity is their simplicity and ease of implementation as well as a potential of ‘super efficiency’. The proposed methods combines the flexibility of a general non stationary model with the efficiency of fast algorithms. In future investigations, it should be possible to develop separation methods which takes full advantage

of both non stationarity and non Gaussianity, at the expense of an increased complexity. Another line of research is an (asymptotic) study of the performance in order to quantify how much information is brought in by considering non stationarity, to give clues regarding the adjustment of the smoothing window for estimating the local variances and for ensuring the stability of on-line algorithms.

APPENDIX: SOME CALCULATIONS

A. Proof of equation (28)

For convenience, we define the matrix

$$\mathbf{M}_t = \sum_{\tau \leq t} \lambda(1-\lambda)^{t-\tau} \boldsymbol{\Sigma}^{-2}(\tau) \mathbf{H} \mathbf{B} \mathbf{X}(\tau) \mathbf{X}(\tau)^T \mathbf{B}^T \quad (41)$$

and also the matrix $\mathbf{U}_t(\mathbf{B})$ by

$$\mathbf{U}_t(\mathbf{B}) = \bar{\mathbf{G}}_t(\mathbf{B}) + \mathbf{I} = \sum_{\tau \leq t} \lambda(1-\lambda)^{t-\tau} \boldsymbol{\Sigma}^{-2}(\tau) \mathbf{B} \mathbf{X}(\tau) \mathbf{X}(\tau)^T \mathbf{B}^T. \quad (42)$$

By definition, matrix $\mathbf{U}_t(\mathbf{B})$ obeys the recursive relation

$$\mathbf{U}_t(\mathbf{B}) = (1-\lambda)\mathbf{U}_{t-1}(\mathbf{B}) + \lambda \boldsymbol{\Sigma}^{-2}(t) \mathbf{B} \mathbf{X}(t) \mathbf{X}(t)^T \mathbf{B}^T. \quad (43)$$

We can also write its first order relative variation in \mathbf{B} by expanding (42):

$$\mathbf{U}_t(\mathbf{B} - \lambda \mathbf{H} \mathbf{B}) = \mathbf{U}_t(\mathbf{B}) - \lambda \mathbf{M}_t - \lambda \mathbf{U}_t(\mathbf{B}) \mathbf{H}^T + O(\lambda^2). \quad (44)$$

Combining (43) and (44) yields

$$\mathbf{U}_t(\mathbf{B} - \lambda \mathbf{H} \mathbf{B}) = (1-\lambda)\mathbf{U}_{t-1}(\mathbf{B}) + \lambda \boldsymbol{\Sigma}^{-2}(t) \mathbf{B} \mathbf{X}(t) \mathbf{X}(t)^T \mathbf{B}^T - \lambda \mathbf{M}_t - \lambda \mathbf{U}_t(\mathbf{B}) \mathbf{H}^T + O(\lambda^2) \quad (45)$$

which can be rearranged as

$$\bar{\mathbf{G}}_t(\mathbf{B} - \lambda \mathbf{H} \mathbf{B}) = \bar{\mathbf{G}}_{t-1}(\mathbf{B}) + \lambda [\boldsymbol{\Sigma}^{-2}(t) \mathbf{B} \mathbf{X}(t) \mathbf{X}(t)^T \mathbf{B}^T - \mathbf{U}_{t-1}(\mathbf{B}) - \mathbf{M}_t - \mathbf{U}_t(\mathbf{B}) \mathbf{H}^T] + O(\lambda^2). \quad (46)$$

Now, if $\bar{\mathbf{G}}_{t-1}(\mathbf{B}) = 0$ then $\mathbf{U}_{t-1}(\mathbf{B}) = \mathbf{I}$ (by definition) and $\mathbf{U}_t(\mathbf{B}) = \mathbf{I} + O(\lambda)$ by (43). Thus, if $\bar{\mathbf{G}}_{t-1}(\mathbf{B}) = 0$, equ. (46) reduces to

$$\bar{\mathbf{G}}_t(\mathbf{B} - \lambda \mathbf{H} \mathbf{B}) = \lambda [\boldsymbol{\Sigma}^{-2}(t) \mathbf{B} \mathbf{X}(t) \mathbf{X}(t)^T \mathbf{B}^T - \mathbf{I} - \mathbf{M}_t - \mathbf{H}^T] + O(\lambda^2) \quad (47)$$

which is the desired result (28).

REFERENCES

- [1] Jean-François Cardoso, "Blind signal separation: statistical principles", *Proceedings of the IEEE. Special issue on blind identification and estimation*, vol. 9, no. 10, pp. 2009–2025, Oct. 1998.
- [2] P. Comon, "Independent component analysis, a new concept ?", *Signal Processing, Elsevier*, vol. 36, no. 3, pp. 287–314, Apr. 1994, Special issue on Higher-Order Statistics.
- [3] Dinh-Tuan Pham and Philippe Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach", *IEEE Tr. SP*, vol. 45, no. 7, pp. 1712–1725, July 1997.
- [4] L. Tong, V.C. Soon, Y.F. Huang, and R. Liu, "AMUSE: a new blind identification algorithm", in *Proc. ISCAS*, 1990.
- [5] Adel Belouchrani, Karim Abed Meraim, Jean-François Cardoso, and Éric Moulines, "A blind source separation technique based on second order statistics", *IEEE Trans. on Sig. Proc.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.
- [6] Antoine Souloumiac, "Blind source detection and separation using second order nonstationarity", in *Proc. ICASSP*, 1995, pp. 1912–1915.
- [7] Michail K. Tsatsanis and Changyeul Kweon, "Source separation using second order statistics: Identifiability conditions and algorithms", in *Proc. 32nd Asilomar Conf. on Signals, Systems, and Computers*. Nov. 1998, pp. 1574–1578, IEEE.

- [8] K. Matsuoka, M. Ohya, and M. Kawamoto, “A neural net for blind separation of nonstationary signals”, *Neural networks*, vol. 8, no. 3, pp. 411–419, 1995.
- [9] J.T. Ngo and N.A. Bhadkamkar, “Adaptive blind separation of audio sources by a physically compact device using second-order statistics”, in *Proc. ICA’99*, Aussois, France, January 11–15, 1999, pp. 257–260.
- [10] Lucas Parra and Clay Spence, “Convulsive blind source separation of non-stationary sources”, *IEEE Trans. on Speech and Audio Processing*, pp. 320–327, may 2000.
- [11] Jean-François Cardoso and Beate Laheld, “Equivariant adaptive source separation”, *IEEE Trans. on Sig. Proc.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.
- [12] S.-I. Amari, *Differential-Geometrical Methods in Statistics*, Number 28 in Lecture Notes in Statistics. Springer-Verlag, 1985.
- [13] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, Wiley series in Telecommunications. John Wiley, 1991.
- [14] Jean-François Cardoso and Antoine Souselmiac, “Blind beamforming for non Gaussian signals”, *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.
- [15] Jean-François Cardoso, “On the performance of orthogonal source separation algorithms”, in *Proc. EUSIPCO*, Edinburgh, Sept. 1994, pp. 776–779.
- [16] Wolfgang Härdle., *Applied nonparametric regression*, Cambridge University press, 1990.
- [17] D.T. Pham, “Joint approximate diagonalization of positive definite Hermitian matrices”, Technical report LMC/IMAG, Laboratoire de Modélisation et de Calcul, BP 53, 38041 Grenoble Cedex 09, France, <http://www-lmc.imag.fr/lmc-sms/Dinh-Tuan.Pham/jadiag/>, Apr. 1999.
- [18] D.T. Pham, “Joint approximate diagonalization of positive definite hermitian matrices”, *SIAM Journal of Matrix Analysis*, 2000, To appear.
- [19] B. N. Flury and W. Gautschi, “An algorithm for the simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly orthogonal form”, *Siam J. of Sci. Stat. Comp.*, vol. 7, no. 1, pp. 169–184, 1986.
- [20] Shun-Ichi Amari, “Natural gradient works efficiently in learning”, *Neural Computation*, vol. 10, pp. 251–276, 1998.
- [21] Lucas Parra and Clay Spence, *Independent Components Analysis: Principles and Practice (Roberts and Everson ed.)*, chapter Separation of non-stationary natural signals, Cambridge University Press, 2000, To appear.
- [22] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, Chapman and Hall, 1995.