Blind Separation of Mixture of Independent Sources Through a Quasi-Maximum Likelihood Approach

Dinh Tuan Pham, *Member, IEEE*, and Philippe Garat

***Abstract*—In this paper, we propose two methods for separating mixtures of independent sources without any precise knowledge of their probability distribution. They are obtained by considering a maximum likelihood (ML) solution corresponding to some given distributions of the sources and relaxing this assumption afterward. The first method is specially adapted to temporally independent non-Gaussian sources and is based on the use of nonlinear separating functions. The second method is specially adapted to correlated sources with distinct spectra and is based on the use of linear separating filters. A theoretical analysis of the performance of the methods has been made. A simple procedure for optimally choosing the separating functions from a given linear space of functions is proposed. Further, in the second method, a simple implementation based on the simultaneous diagonalization of two symmetric matrices is provided. Finally, some numerical and simulation results are given, illustrating the performance of the method and the good agreement between the experiments and the theory.**

***Index Terms*— Higher order moments, independence, linear filters, maximum likelihood, separation of sources.**

I. Introduction

THIS paper deals with the problem of blind separation of a mixture of independent sources without any precise knowledge of their probability distribution. This problem has important applications in signal processing (e.g., speech analysis, radar and sonar processing, etc.). Many methods have been proposed in the literature ([2]–[7], [9]–[11], [13], [20], etc.); most of them are based on higher order moment statistics. Our approach is to consider the maximum likelihood (ML) solution corresponding to a given hypothetical model of sources, which we shall relax later. For temporally independent (i.e., white) sources, it leads to more general statistics nonlinear functions of the data. A similar algorithm derived in an *ad hoc* manner has appeared in Jutten and Hérault [9], but these authors use only two separating functions, whereas we use as many as the number of sources. Further, instead of choosing such functions *a priori* as these authors, we also have made a performance study and derived a method for choosing them optimally in a given linear space of functions. Our results have appeared in a condensed form in [14] and [15]. We also propose an adaptive method in order to follow changes of the mixture weights. The ML principle also has been applied in [7] with the source density approximated by a Gram–Charlier expansion based on cumulants up to the fourth order. This leads to a solution that is essentially fourth-order cumulant based, whereas our method uses general nonlinear statistics.

In practice, the sources are rarely white, but our separation procedure still works since it only exploits the independence between different sources. However, one may expect that a method exploiting the temporal dependency of the sources as well might perform better. Therefore, we introduce another method that is also likelihood based but taking into account the correlation structure of the sources (which are assumed to be stationary). This leads to a set of linear *separating filters* instead of nonlinear *separating functions*. Note that this new procedure exploits only the second-order lagged covariance structure of the data and is thus able to separate mixture of *Gaussian* sources, provided only that they have distinct spectra. Further, since it is based on second-order statistics, it requires very little computation. In particular, in the case of two sources, our procedure reduces to the simultaneous diagonalization of two covariance matrices: No iterative computation is necessary. This simple separation procedure also works in the general case, but its performance will not be optimal. More generally, we have derived an expression for the asymptotic covariance matrix of our estimators and for their lowest asymptotic bound achievable.

In the following, the observation record will be denoted by $\mathbf{X}(1), \cdots, \mathbf{X}(T)$, where each $\mathbf{X}(t)$ is a random vector of K' components $X_1(t), \cdots, X_{K'}(t)$ corresponding to K' observation channels. The underlying assumption is that each $X_i(t)$ is a linear combination of K independent sources: $\mathbf{X}(t) = \mathbf{A}\mathbf{S}(t)$, where $\mathbf{A}$ is some unknown matrix, and $\{\mathbf{S}(t), t = 1, 2, \cdots\}$ is a stationary sequence of random vectors with K components $S_1(t), \cdots, S_K(t)$ and the sequences $\{S_k(1), S_k(2), \cdots\}$, $k = 1, \cdots, K$ being mutually independent. The source separation problem is to reconstruct the "sources" $S_i(t)$ from the observations. The separation is called blind because it is based only on the independence between the sources without any probability model for them.

To simplify the problem, we shall, in the sequel, restrict ourselves to the case where the number K sources equals the number K' sensors. Note that when $K < K'$, the observation vector $\mathbf{X}(t)$ lies on a linear subspace of dimension K of $\mathbb{R}^{K'}$; hence, by a change of basis, its last $K' - K$ components can be made to vanish, and one is led back to the case $K = K'$. In practice, the situation can be more complicated as the observations may be corrupted with additive noises, in which is case, the setup is similar to the factor analysis

Manuscript received September 20, 1995; revised February 24, 1997. The associate editor coordinating the review of this paper and approving it for publication was Dr. Zhi Ding.

The authors are with the Laboratory of Modeling and Computation, IMAG, Centre National de Recherche Scientifique, University of Grenoble, Grenoble, France.

Publisher Item Identifier S 1053-587X(97)04951-9.

considered in the psychometric literature (see, for example, [17]). However, in factor analysis, both the factor scores (i.e., sources) and noises are assumed Gaussian, and the adopted method is second-order based and yields the factor loading (the matrix $\mathbf{A}$) only up to an orthogonal rotation. Another feature that distinguishes factor analysis with our problem is that the former is an identification and not a separation problem. With K not too high with respect to K',[1] one can identify the factor loading and the noise variances, but the factor scores (i.e., sources) cannot be reconstructed: The best one can do is to estimate their conditional expectations given the observed channels, which would always contain some contamination from the noises.

II. The ML Approach for "White" Sources

A. Derivation of the Separation Procedure

In this section, we will assume that the sources are white in the sense that the $S_i(t)$ at different times t are independent and identically distributed. This assumption may be unrealistic, but it will be used *only as a working assumption* in order to write down the likelihood and derive a separation procedure. We shall see later that the method can be justified without it.

To write down the likelihood, one needs the density functions of the sources, which is unknown in a blind context. For this reason and the inherent scaling ambiguity of the problem (see [2] and [19]), we shall assume for the moment that these densities are known up to a scale factor. Thus, let the source $S_i(t)$ has density $f_i(\cdot/\sigma_i)/\sigma_i$ with f_i known and σ_i unknown. Then $\mathbf{S}(t)$ has density $\prod_{i=1}^{K} f_i[S_i(t)/\sigma_i]/\sigma_i$. Since $\mathbf{X}(t)$ is related to $\mathbf{S}(t)$ through $\mathbf{X}(t) = \mathbf{A}\mathbf{S}(t)$, the logarithm of the density of the data can be written as (assuming $\mathbf{A}$ invertible):

$$L_T = T\left\{\sum_{i=1}^{K} \hat{E} \ln\left[\frac{n}{\sigma_i} f\left(\frac{\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{X}}{\sigma_i}\right)\right] - \ln \mid \det \mathbf{A}\mid\right\}$$

where

- $\hat{E}$ time average operator $\hat{E}[g(\mathbf{X})] = \{g[\mathbf{X}(1)] + \cdots + g[\mathbf{X}(T)]\}/T$;
- $\mathbf{e}_i$ ith column of the identity matrix (of order K);
- ^{T} transpose.

The above expression, as a function of the unknown parameter, is the log likelihood function, and the ML method consists of maximizing it. To this end, one may equate the partial derivatives of L_T to zero, but we find it more convenient to work with the differential instead. We note that $d\mathbf{A}^{-1} = -\mathbf{A}^{-1}\,d\mathbf{A}\mathbf{A}^{-1}$, $d \ln \mid \det \mathbf{A}\mid = tr(\mathbf{A}^{-1}\,d\mathbf{A})$, where ($tr$ denotes the trace). Hence, under the assumption that f_i is differentiable and putting $\psi_i^\dagger = -(\ln f_i)'$, where prime $'$ denotes the derivative, we get

$$\begin{aligned} T^{-1}\,dL_T &= \sum_{i=1}^{K} \hat{E}\left[\psi_i^\dagger\left(\frac{\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{X}}{\sigma_i}\right)\frac{\mathbf{e}_i^T\mathbf{A}^{-1}\,d\mathbf{A}\mathbf{A}^{-1}\mathbf{X}}{\sigma_i}\right] \\ &\quad - \operatorname{tr}(\mathbf{A}^{-1}\,d\mathbf{A}) \\ &\quad + \sum_{i=1}^{K} \hat{E}\left[\psi_i^\dagger\left(\frac{\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{X}}{\sigma_i}\right)\frac{\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{X}}{\sigma_i^2} - \frac{1}{\sigma_i}\right] d\sigma_i. \end{aligned}$$

Denoting by ∂_{ij} the general element of $\mathbf{A}^{-1}\,d\mathbf{A}$, the first two terms of the above right-hand side becomes

$$\sum_{i=1}^{K}\sum_{j=1}^{K} \hat{E}\left[\psi_i^\dagger\left(\frac{\mathbf{e}_i^T\mathbf{A}^{-1}\mathbf{X}}{\sigma_i}\right)\frac{\mathbf{e}_j^T\mathbf{A}^{-1}\mathbf{X}}{\sigma_i}\right]\partial_{ij} - \sum_{i=1}^{K}\partial_{ii}.$$

At the ML estimators $\hat{\mathbf{A}}$ and $\hat{\sigma}_i$ of $\mathbf{A}$ and σ_i, dL_T must vanish for all infinitesimal increments $d\mathbf{A}$ (or equivalently ∂_{ij}) and $d\sigma_i$. This yields the estimating equations:

$$\hat{E}\left[\psi_i^\dagger\left(\frac{\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X}}{\hat{\sigma}_i}\right)\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}\right] = 0, \quad i \neq j = 1, \cdots, K$$

$$\hat{\sigma}_i = \hat{E}\left[\psi_i^\dagger\left(\frac{\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X}}{\hat{\sigma}_i}\right)\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X}\right], \quad i = 1, \cdots, K.$$

Note that the parameter set $\{\mathbf{A}, \sigma_1, \cdots, \sigma_K\}$ is redundant: Post-multiplying $\mathbf{A}$ by a diagonal matrix and dividing the σ_i by its corresponding diagonal elements would not change the likelihood. Thus, $\mathbf{A}$ can only be estimated up to the post-multiplication by a diagonal matrix, possibly permuted, which merely reflects the scaling and permutation ambiguities of the problem (see [19]). The first $K(K-1)$ equations above in fact determine $\hat{\mathbf{A}}$ up to scaling, and once a scaling convention is adopted to make it unique, the last K equations merely serve to estimate the σ_i. Since we are not interested in their values, which would depend on the scaling convention, we just drop these equations. On the other hand, the densities of the sources being unknown in the blind separation context, instead of $\psi_i^\dagger$, one has to take some *a priori* function ψ_i, which will be also made to absorb the factor $\hat{\sigma}_i$. Thus, the proposed separation procedure consists of solving, with respect to $\hat{\mathbf{A}}$, the following system of estimating equations:

$$\hat{E}[\psi_i(\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X})\mathbf{e}_j^{\mathrm{T}}\hat{\mathbf{A}}^{-1}\mathbf{X}] = 0, \qquad i \neq j = 1, \cdots, K. \tag{2.1}$$

There are just enough equations, namely $K(K-1)$, to solve for $\hat{\mathbf{A}}$ up to a scaling factor for each of its columns. In applications, some scaling convention must be adopted to make this matrix unique.

Note that by choosing the ψ_i *a priori*, the above procedure is *no longer ML*. Nevertheless, the procedure is still justified if the sources are centered and possess the stationarity and ergodicity property, as will be assumed in this paper. Indeed, under these assumptions, (2.1) is asymptotically satisfied with $\hat{\mathbf{A}}$ replaced by the true mixing matrix up to a scaling and a permutation since for such a matrix, the left hand sides of (2.1) converge to the expectations of $\psi_i[S_i(t)/\tilde{\sigma}_i]S_j(t)$, $i \neq j$ with

[1] The number of parameters in the factor model is $K'K - K(K-1)/2 + K'$ for the factor loading $\mathbf{A}$ (taking into account of the indeterminacy associated with rotation) and the noise variances. Since the identification is based on the $K'(K'+1)/2$ elements of the sample covariance matrix of the observations, one must have $K'K - K(K-1)/2 + K' \leq K'(K'+1)/2$ or $K \leq K' + 1/2 - \sqrt{2K'+1/4}$.

$\tilde{\sigma}_i$ being some scaling factors, and these expectations vanish as $S_j(t)$ has zero mean and is independent of $S_i(t)$, $i \neq j$.

A similar *ad hoc* procedure has been proposed by Jutten and Hérault [9]. It is an adaptive procedure but can be easily changed into a batch processing one, which amounts to solving

$$\hat{E}[\psi(\mathbf{e}_i^T \hat{\mathbf{A}}^{-1}\mathbf{X})\varphi(\mathbf{e}_j^T \hat{\mathbf{A}}^{-1}\mathbf{X})] = 0, \qquad i \neq j = 1, \cdots, K$$

where ψ and φ are two odd functions. Our approach leads, however, to taking φ to be the identity function and using different functions ψ (which are not restricted to be odd) for each source.

B. Asymptotic Properties of the Estimator

It should be noted that a separation of source procedure does not provide an estimator of the mixing matrix but a whole set of estimators, related to each other by post multiplication by a permuted diagonal matrix. Thus, a permutation and a scaling convention must be adopted to get an unique estimator $\hat{\mathbf{A}}$. A natural permutation convention is to permute the columns of $\hat{\mathbf{A}}$ such that, denoting by $\hat{A}_{ij}$, the general element of $\hat{\mathbf{A}}$, $|\hat{A}_{11} \cdots \hat{A}_{KK}|$ is largest among the $|\hat{A}_{1p_1} \cdots \hat{A}_{Kp_K}|$ with $p_1, \cdots, p_K$ running through all permutations of $1, \cdots, K$. As for scaling convention, a simple one is to make $\hat{\mathbf{A}}$ having unit diagonal elements, but other conventions may be used (such as making the sample covariance matrix of $\hat{\mathbf{A}}^{-1}\mathbf{X}(t)$ having unit diagonal elements). Note that because of such scaling, the procedure is not truly ML even if ψ_i happens to equal $\psi^* = -(\ln f_i^*)'$, where f_i^* denotes the true density of the ith source.

Because of the permutation and scaling ambiguities, a "good" separation procedure needs not produce an $\hat{\mathbf{A}}$ close to the true $\mathbf{A}$ but only that $\hat{\mathbf{A}}^{-1}\mathbf{A}$ close to a permuted diagonal matrix. However, if we *permute and scale both* $\hat{\mathbf{A}}$ and $\mathbf{A}$ *by the same convention*, we may expect that $\boldsymbol{\delta} = \mathbf{I} - \hat{\mathbf{A}}^{-1}\mathbf{A}$ is small. We can then make a Taylor expansion of (2.1) around $\mathbf{A}$. From the relation $\hat{\mathbf{A}}^{-1} = \mathbf{A}^{-1} - \boldsymbol{\delta}\mathbf{A}^{-1}$, we have

$$\mathbf{e}_i^T \hat{\mathbf{A}}^{-1}\mathbf{X}(t) = S_i(t) - \sum_{j=1}^{K} \delta_{ij} S_j(t) \tag{2.2}$$

where δ_{ij} is the general element of $\boldsymbol{\delta}$, and $S_i(t) = \mathbf{e}_i^T \mathbf{A}^{-1}\mathbf{X}(t)$ is the ith source (possibly scaled and permuted as a result of the permutation and scaling convention applied to $\mathbf{A}$). Hence, by Taylor expansion, the left-hand side of (2.1) approximatively equals

$$\hat{E}[\psi_i(S_i)S_j] - \sum_{k=1}^{K} \{\hat{E}[\psi_i'(S_i)S_jS_k]\delta_{ik} + \hat{E}[\psi_i(S_i)S_k]\delta_{jk}\}.$$

As $T \to \infty$, $\hat{E}[\psi_i'(S_i)S_jS_k]$ and $\hat{E}[\psi_i(S_i)S_k]$ converge to the expectation of $\psi_i'(S_i)S_jS_k$ and $\psi_i(S_i)S_k$, which vanish unless $k = j$ or $k = i$. Thus, (2.1) becomes

$$\begin{aligned} &E[\psi_i'(S_i)S_j^2]\delta_{ij} + E[\psi_i(S_i)S_i]\delta_{ji} \\ &\quad \approx \hat{E}[\psi_i(S_i)S_j], \qquad i \neq j = 1, \cdots, K \end{aligned} \tag{2.1'}$$

where E is the expectation operator. Write $\boldsymbol{\delta}$ in the form of a $K(K-1)$-vector $\boldsymbol{\Delta} = (\delta_{12}\ \delta_{21} \cdots)^T$, and let $\boldsymbol{\Psi} = (\Psi_{12}\ \Psi_{21} \cdots)^T$, where $\Psi_{ij} = \psi_i(S_i)S_j$. The last equation can be written as $\mathbf{H}\boldsymbol{\Delta} \approx \hat{E}(\boldsymbol{\Psi})$, where $\mathbf{H}$ is the block diagonal matrix with blocks

$$\mathbf{H}_{(ij)} = \begin{bmatrix} E[\psi_i'(S_i)]E(S_j^2) & E[\psi_i(S_i)S_i] \\ E[\psi_j(S_j)S_j] & E[\psi_j'(S_j)]E(S_i^2) \end{bmatrix} \tag{2.3}$$

indexed by the pairs (i, j), $i \neq j = 1, \cdots, K$. On the other hand, by the central limit theorem, $\hat{E}(\boldsymbol{\Psi})$ is asymptotically Gaussian with mean zero and covariance matrix $\mathbf{G}/T$, where $\mathbf{G} = E(\boldsymbol{\Psi}\boldsymbol{\Psi}^T)$ if the $\mathbf{S}(t)$ are temporally independent [12]. The last assumption is made here only to simplify the computation. The central limit theorem can be applied for ergodic sequences satisfying mild conditions [16], but then, $\mathbf{G}$ will be too complicated for further analysis. Thus, $\boldsymbol{\Delta}$ is asymptotically Gaussian with covariance matrix $\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-T}/T$, provided that $\mathbf{H}$ is invertible.[2] Here, and in the sequel, the notation $\mathbf{H}^{-T}$ is used as short hand for $(\mathbf{H}^{-1})^T$. Note that $\mathbf{G}$ is also block diagonal if the random variables $\psi_i[S_i(t)]$ have zero mean, which is the case, for example, if the $S_i(t)$ have symmetric distributions and the ψ_i are odd functions. In this case, the random vectors $(\delta_{ij}\ \delta_{ji})^T$, for different pairs (i, j), are asymptotically independent with covariance matrix $\mathbf{H}_{(ij)}^{-1}\mathbf{G}_{(ij)}\mathbf{H}_{(ij)}^{-T}/T$, where

$$\mathbf{G}_{(ij)} = \begin{bmatrix} E[\psi_i^2(S_i)]E(S_j^2) & E[\psi_i(S_i)S_i]E[\psi_j(S_j)S_j] \\ E[\psi_i(S_i)S_i]E[\psi_j(S_j)S_j] & E[\psi_j^2(S_j)]E(S_i^2) \end{bmatrix}. \tag{2.4}$$

We have thus derived the asymptotic distribution of the off-diagonal terms of the matrix $\boldsymbol{\delta}$. This distribution does not depend on the unknown mixing matrix $\mathbf{A}$ but only on the sources characteristics, which is one reason why we consider $\boldsymbol{\delta}$ and not $\hat{\mathbf{A}} - \mathbf{A}$. Another reason is that the δ_{ij} are directly related to a set of natural performance indexes called contamination coefficients, which completely characterize the separation procedure. The right-hand side of (2.2) represents the reconstructed ith source, which can be written as $(1 - \delta_{ii})(S_i - \sum_{i \neq j} c_{ij}S_j)$, where $c_{ij} = \delta_{ij}/(1 - \delta_{ii})$. Thus, $c_{ij}S_j$ represents the contamination of the jth source in the reconstructed ith source and, therefore, c_{ij} will be called the contamination coefficient. Since $\delta_{ii} \to 0$ as $T \to \infty$ (as both $\hat{\mathbf{A}}$ and $\mathbf{A}$ are permuted and scaled by the same convention), the c_{ij}, $i \neq j$ *have the same asymptotic distribution* as the δ_{ij}. Note that the sources S_i may be scaled and have different variances (powers). Therefore, it is preferable to work with the relative contamination coefficient $c_{ij}\sigma_j/\sigma_i$, where σ_i^2 denotes the variance of the ith source.

Using the fact that the ML estimator is asymptotically efficient (under mild conditions), one could obtain that the asymptotic covariance matrix of our estimator is minimum when ψ is chosen proportional to ψ^*. However, this argument is not quite rigorous as we have dropped the K "scaling" equations in the likelihood equations and introduced a scaling

[2] As can be seen from (2.5) below and the fact that ψ_i^* is the identity function for Gaussian sources, this is not the case if more than two sources are Gaussian.

convention leading to the restriction to the off-diagonal elements of $\boldsymbol{\delta}$. Therefore, we give here a brief direct proof of the result. First note that by integration by parts

$$E[g(S_i)\psi_i^*(S_i)] = E[g'(S_i)] \tag{2.5}$$

for any differentiable function g such that gf_i^* vanished at infinity, provided that f_i^* is differentiable so that ψ_i^* is well defined. (This result is also useful later,) Then, it can be checked that

$$E\left(\begin{bmatrix}\boldsymbol{\Psi}\\ \boldsymbol{\Psi}^*\end{bmatrix}[\boldsymbol{\Psi}^T\,(\boldsymbol{\Psi}^*)^T]\right) = \begin{bmatrix}\mathbf{G} & \mathbf{H}\\ \mathbf{H}^T & \mathbf{J}\end{bmatrix}$$

where $\mathbf{J} = E[\boldsymbol{\Psi}^*(\boldsymbol{\Psi}^*)^T]$. From the non-negativity of the matrix in the above left-hand side, $\mathbf{G} - \mathbf{H}\mathbf{J}^{-1}\mathbf{H}^T \geq \mathbf{0}$, yielding $\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-T}/T \geq \mathbf{J}^{-1}/T$, and it can be easily seen that equality is attained when ψ_i is taken proportional to ψ_i^*, hence, the announced result. Note that the matrix $\mathbf{J}$ is also block diagonal as $\mathbf{H}$ since $E[\psi_i^*(S_i)] = 0$.

We now derive explicitly the asymptotic covariance matrix of $(\delta_{12}\ \delta_{21}\ \cdots)^T$ in terms of the source characteristics. Since there exists a decoupling between different pairs of channels (under mild assumption, see above), we need only to compute for each pair (i, j) the asymptotic covariance matrix $\mathbf{H}_{(ij)}^{-1}\mathbf{G}_{(ij)}\mathbf{H}_{(ij)}^{-T}/T$ of $(\delta_{ij}\ \ \delta_{ji})^T$. It is also more interesting to look at the asymptotic covariance matrix of the relative contamination coefficients, which is the same as that of $\delta_{ij}\sigma_j/\sigma_i$. To this end, we introduce the dimensionless parameters:

$$\rho_i = \frac{E[\psi_i(S_i)S_i]}{\sqrt{E[\psi_i^2(S_i)]}\sigma_i} = \text{corr}\,\{\psi_i(S_i),\, S_i\}$$
$$\lambda_i = \frac{E[\psi_i(S_i)S_i]}{E[\psi_i'(S_i)]\sigma_i^2}.$$

Then, denoting by $\mathbf{D}_{(ij)}$ the diagonal matrix with diagonal elements $\sqrt{E[\psi_i^2(S_i)]}\sigma_j$, $\sqrt{E[\psi_j^2(S_j)]}\sigma_i$, we have

$$\mathbf{G}_{(ij)} = \mathbf{D}_{(ij)}\begin{bmatrix}1 & \rho_i\rho_j\\ \rho_i\rho_j & 1\end{bmatrix}\mathbf{D}_{(ij)}$$
$$\mathbf{H}_{(ij)} = \mathbf{D}_{(ij)}\begin{bmatrix}\frac{\rho_i}{\lambda_i} & \rho_i\\ \rho_j & \frac{\rho_j}{\lambda_j}\end{bmatrix}\begin{bmatrix}\frac{\sigma_j}{\sigma_i} & 0\\ 0 & \frac{\sigma_i}{\sigma_j}\end{bmatrix}.$$

Thus, the asymptotic covariance matrix of $(\delta_{ij}\sigma_j/\sigma_i\delta_{ji}\ \sigma_i/\sigma_i)^T$ can be seen to be

$$\frac{1}{T}\begin{bmatrix}\frac{\rho_i}{\lambda_i} & \rho_i\\ \rho_j & \frac{\rho_j}{\lambda_j}\end{bmatrix}^{-1}\begin{bmatrix}1 & \rho_i\rho_j\\ \rho_i\rho_j & 1\end{bmatrix}\begin{bmatrix}\frac{\rho_i}{\lambda_i} & \rho_j\\ \rho_i & \frac{\rho_j}{\lambda_j}\end{bmatrix}^{-1}.$$

Simple algebra then yields that the above matrix equals (2.5a), shown at the bottom of the page. The diagonal elements of this matrix represent the mean square relative contaminations of the jth source to the reconstructed ith source and vice versa, and the off-diagonal element represents their covariance. These indexes *depend only on the dimensionless parameters* $\rho_i, \rho_j, \lambda_i,$ *and* λ_j. The same is true for the asymptotic covariance between δ_{ij} and δ_{ji}.

Significant simplifications of the above results arise when $\lambda_i = \rho_i^2$. This happens when ψ_i is chosen proportional to the orthogonal projection of ψ_i^* onto a linear space of functions containing the identity function (in particular ψ_i^* itself). Indeed, in such a case $E[\psi_i(S_i)S_i] = \alpha E[\psi_i^*(S_i)S_i]$, $E[\psi_i^2(S_i)] = \alpha E[\psi_i(S_i)\psi_i^*(S_i)]$ for some real number α. From this and the equalities $E[\psi_i'(S_i)] = E[\psi_i(S_i)\psi_i^*(S_i)]$ and $E[\psi_i^*(S_i)S_i] = 1$, which are consequences of (2.5), the definition of λ_i can be seen indeed to reduce to ρ_i^2. In the case where $\lambda_i = \rho_i^2$

$$\frac{\lambda_i^2\lambda_j^2}{\rho_i^2\rho_j^2}\left(\frac{\rho_j^2}{\lambda_j^2} + \rho_i^2 - 2\frac{\rho_i^2\rho_j^2}{\lambda_j}\right) = \rho_i^2\rho_j^2\left(\frac{1}{\rho_j^2} + \rho_i^2 - 2\rho_i^2\right)$$
$$= \rho_i^2(1-\rho_i^2\rho_j^2)$$
$$\frac{\lambda_i^2\lambda_j^2}{\rho_i^2\rho_j^2}\left[\frac{\rho_i^2\rho_j^2(1+\lambda_i\lambda_j)}{\lambda_i\lambda_j} - \left(\frac{\rho_j^2}{\lambda_j} + \frac{\rho_i^2}{\lambda_i}\right)\right]$$
$$= \rho_i^2\rho_j^2(1+\rho_i^2\rho_j^2-2) = -\rho_i^2\rho_j^2(1-\rho_i^2\rho_j^2).$$

Hence, the asymptotic covariance matrix of $(\delta_{ij}\sigma_j/\sigma_i\ \ \delta_{ji}\sigma_i/\sigma_j)^T$ reduces to

$$\frac{1}{T(1-\rho_i^2\rho_j^2)}\begin{bmatrix}\rho_i^2 & -\rho_i^2\rho_j^2\\ -\rho_i^2\rho_j^2 & \rho_j^2\end{bmatrix} = \frac{1}{T}\begin{bmatrix}\rho_i^{-2} & 1\\ 1 & \rho_j^{-2}\end{bmatrix}^{-1}.$$

The last equality shows that (in the case where $\lambda_i = \rho_i^2$) the asymptotic covariance matrix of the estimator increases with ρ_i. Further, since the lower bound for the asymptotic covariance matrix of the estimator is attained when ψ_i is chosen proportional to ψ_i^*, it is given by any one of the two sides of the last equality with ρ_i replaced by $\rho_i^* = \text{corr}\,\{\psi_i^*(S_i),\, S_i\}$.

A different simplification is obtained in the case where the signals S_i and S_j have the same distribution. Then, $\sigma_i = \sigma_j = \sigma$ and $\lambda_i = \lambda_j = \lambda$, and the above asymptotic covariance matrix of $(\delta_{ij}\ \delta_{ji})^T$ reduces to

$$\frac{\lambda^2}{(1-\lambda^2)^2\rho^2}\begin{bmatrix}1+\lambda^2-2\lambda\rho^2 & \rho^2(1+\lambda^2)-2\lambda\\ \rho^2(1+\lambda^2)-2\lambda & 1+\lambda^2-2\lambda\rho^2\end{bmatrix}.$$

$$\frac{\lambda_i^2\lambda_j^2}{(1-\lambda_i\lambda_j)^2\rho_i^2\rho_j^2}\begin{bmatrix}\frac{\rho_j^2}{\lambda_j^2} + \rho_i^2 - 2\frac{\rho_i^2\rho_j^2}{\lambda_j} & \frac{\rho_i^2\rho_j^2(1+\lambda_i\lambda_j)}{\lambda_i\lambda_j} - \left(\frac{\rho_j^2}{\lambda_j} + \frac{\rho_i^2}{\lambda_i}\right)\\ \frac{\rho_i^2\rho_j^2(1+\lambda_i\lambda_j)}{\lambda_i\lambda_j} - \left(\frac{\rho_j^2}{\lambda_j} + \frac{\rho_i^2}{\lambda_i}\right) & \frac{\rho_i^2}{\lambda_i^2} + \rho_j^2 - 2\frac{\rho_j^2\rho_i^2}{\lambda_i}\end{bmatrix}. \tag{2.5a}$$

If, moreover, $\lambda = \rho^2$, by the same computation as above, the asymptotic covariance matrix of $(\delta_{ij}\ \delta_{ji})^T$ further reduces to

$$\frac{\rho^2}{T(1-\rho^4)}\begin{bmatrix} 1 & -\rho^2 \\ -\rho^2 & 1 \end{bmatrix} = \frac{1}{T}\begin{bmatrix} \rho^{-2} & 1 \\ 1 & \rho^{-2} \end{bmatrix}^{-1}.$$

As before, this matrix, with $\rho = \rho^* = \text{corr}\ \{\psi_i^*(S_i), S_i\}$ (which does not depend on i), constitutes the lower bound for the asymptotic covariance matrix of the estimator.

C. Construction of the Estimator

An algorithm for finding the solution of (2.1) similar to the Fisher's scoring technique for maximizing likelihood and based on the Newton–Raphson iteration (see, for example, [16]), is now described. By expanding (2.1) around an initial estimate $\tilde{\mathbf{A}}$ of $\mathbf{A}$, using the same computation as in Section II-B, the next step estimate, which is denoted by $\tilde{\mathbf{A}}(\mathbf{I}-\tilde{\boldsymbol{\delta}})^{-1}$, can be constructed as the solution of the linear system

$$\sum_{k=1}^{K}\{\hat{E}[\psi_i'(\tilde{S}_i)\tilde{S}_j\tilde{S}_k]\tilde{\delta}_{ik} + \hat{E}[\psi_i(\tilde{S}_i)\tilde{S}_k]\tilde{\delta}_{jk}\} = \hat{E}[\psi_i(\tilde{S}_i)\tilde{S}_j], \qquad i \neq j = 1, \cdots, K$$

where $\tilde{S}_i = \mathbf{e}_i^T\tilde{\mathbf{A}}^{-1}\mathbf{X}$, and $\tilde{\delta}_{ij}$ is the general element of $\tilde{\boldsymbol{\delta}}$. Since $\tilde{S}_i(t)$ is close to $S_i(t)$, one may approximate $\hat{E}[\psi_i'(\tilde{S}_i)\tilde{S}_j\tilde{S}_k]$ by $E[\psi_i'(S_i)S_jS_k]$, which vanishes if $k \neq j$, and $\hat{E}[\psi_i(\tilde{S}_i)\tilde{S}_k]$ by $E[\psi_i(S_i)S_k]$, which vanishes if $k \neq i$. Thus, the above system of equations may be replaced by

$$\hat{E}[\psi_i'(\tilde{S}_i)\tilde{S}_j^2]\tilde{\delta}_{ij} + \hat{E}[\psi_i(\tilde{S}_i)\tilde{S}_i]\tilde{\delta}_{ji} = \hat{E}[\psi_i(\tilde{S}_i)\tilde{S}_j], \qquad i \neq j = 1, \cdots K.$$

This system defines the off-diagonal elements of $\boldsymbol{\delta}$; the diagonal ones, in fact, can be arbitrary as long as they are "small" since they enter the equations through the multiplication with $\hat{E}[\psi_i'(\tilde{S}_i)\tilde{S}_i\tilde{S}_j]$, $j \neq i$, which almost vanish. We recommend that they be put to zero and apply further a scaling convention to the new estimate, which amounts to rescaling the rows of $\mathbf{I}-\tilde{\boldsymbol{\delta}}$. This prevents the new estimator from becoming too large or too small. The procedure may then be iterated until convergence.

The above algorithm can be easily modified for use in an adaptive context, in which the estimate is updated each time a new observation is available. Let $\tilde{\mathbf{A}}(t-1)$ be an estimate of $\mathbf{A}$ at time $t-1$. The application of the above iteration leads to the new estimate at time t: $\tilde{\mathbf{A}}(t) = \tilde{\mathbf{A}}(t-1)[\mathbf{I}-\tilde{\boldsymbol{\delta}}(t)]^{-1}$, where $\tilde{\boldsymbol{\delta}}(t)$ is the matrix with the element $\tilde{\delta}_{ij}(t)$, $i \neq j$ as a solution of

$$\left\{\sum_{s=1}^{t}\psi_i'[\tilde{S}_i(s)]\tilde{S}_j^2(s)\right\}\tilde{\delta}_{ij}(t) + \left\{\sum_{s=1}^{t}\psi_i[\tilde{S}_i(s)]\tilde{S}_i(s)\right\}\tilde{\delta}_{ji}(t) = \sum_{s=1}^{t}\psi_i[\tilde{S}_i(s)]\tilde{S}_j(s).$$

Note that $\tilde{S}_i(.)$ is now computed from $\tilde{\mathbf{A}}(t-1)$ and that the estimate $\tilde{\mathbf{A}}(t-1)$ has been constructed in order to satisfy, at least approximately, the likelihood equation based on the data $\mathbf{X}(1), \cdots, \mathbf{X}(t-1)$. Hence, $\sum_{s=1}^{t-1}\psi_i[\tilde{S}_i(s)]\tilde{S}_j(s)$ should be close to 0. By making this approximation, the above right-hand side reduces to $\psi_i[\tilde{S}_i(t)]\tilde{S}_j(t)$, yielding the algorithm

$$\begin{aligned}
\tilde{S}_i(t) &= \mathbf{e}_i^T\tilde{\mathbf{A}}(t-1)^{-1}\mathbf{X}(t),\\
\Omega_{ii}(t) &= \lambda\Omega_{ii}(t-1) + \psi_i[\tilde{S}_i(t)]\tilde{S}_i(t),\\
\Omega_{ij}(t) &= \lambda\Omega_{ij}(t-1) + \psi_i'[\tilde{S}_i(t)]\tilde{S}_j^2(t), \qquad i \neq j\\
\tilde{\delta}_{ij}\ \tilde{\delta}_{ji}&\ \text{solution of: } \Omega_{ij}(t)\tilde{\delta}_{ij}(t) + \Omega_{ii}(t)\tilde{\delta}_{ji}(t)\\
&= \psi_i[\tilde{S}_i(t)]\tilde{S}_j(t), \qquad i \neq j
\end{aligned}$$

where $\lambda \in (0, 1]$ is a forgetting factor introduced to track the possible evolution of the mixing matrix $\mathbf{A}$.

D. Choice of Separating Function

Since the optimal choice for ψ_i depends on the unknown distribution of the source S_i, which is not easily estimated, we shall restrict the choice to a linear space of functions spanned by a given basis $\phi_1, \cdots, \phi_N$: $\psi_i = \sum_{n=1}^{N} c_{in}\phi_n$ for some coefficients c_{in}. The goal is to minimize the asymptotic covariance matrix of the estimator or, equivalently, to maximize $\mathbf{H}_{(ij)}^T\mathbf{G}_{(ij)}^{-1}\mathbf{H}_{(ij)}$, where $\mathbf{H}_{(ij)}$ and $\mathbf{G}_{(ij)}$ are defined in Section II-B. The following result shows that this can be done quite easily through a projection.

Proposition 2.1: Assume that one of the basis function ϕ_n is the identity function. Then, the matrices $\mathbf{H}_{(ij)}^T\mathbf{G}_{(ij)}^{-1}\mathbf{H}_{(ij)}$ are maximized among all choices of $\psi_i = \sum_{n=1}^{N} c_{in}\phi_n$ when ψ_i is chosen proportional to the orthogonal projection of ψ_i^* onto the space spanned by $\phi_1, \cdots, \phi_N$. Explicitly, the optimal choice corresponds to taking $c_{in} = c_{in}^*$, $n = 1, \cdots, N$, which is the solution of the linear system

$$\sum_{n=1}^{N} c_{in}^* E[\phi_n(S_i)\phi_m(S_i)] = E[\psi_i^*(S_i)\phi_m(S_i)], \qquad m = 1, \cdots, N. \quad (2.6)$$

Proof: By (2.3)–(2.5), the matrices $\mathbf{H}_{(ij)}^T$ and $\mathbf{G}_{(ij)}$ can be written as $E[\boldsymbol{\Psi}_{(ij)}^*\boldsymbol{\Psi}_{(ij)}^T]$ and $E[\boldsymbol{\Psi}_{(ij)}\boldsymbol{\Psi}_{(ij)}^T]$, where $\boldsymbol{\Psi}_{(ij)} = [\psi_i(S_i)S_j\,\psi_j(S_j)S_i]^T$ and $\boldsymbol{\Psi}_{(ij)}^*$ is defined similarly with ψ_i^* in place of ψ_i. By assumption, one of the ϕ_i, which we may take to be ϕ_1, is the identity function. Thus, $\boldsymbol{\Psi}_{(ij)} = \mathbf{C}_{(ij)}\boldsymbol{\Phi}_{(ij)}$, where

$$\mathbf{C}_{(ij)} = \begin{bmatrix} c_{i1} & c_{i2} & \cdots & c_{iN} & 0 & \cdots & 0 \\ c_{j1} & 0 & \cdots & 0 & c_{j2} & \cdots & c_{jN} \end{bmatrix}$$

with $\boldsymbol{\Phi}_{(ij)} = [S_iS_j\ \phi_2(S_i)S_j\ \cdots\phi_N(S_i)S_j\,\phi_2(S_j)S_i\cdots\phi_N(S_j)S_i]^T$. On the other hand, the orthogonal projection $\boldsymbol{\Psi}_{ij}^{\dagger}$ of $\boldsymbol{\Psi}_{ij}^*$ onto the linear space spanned by the components of $\boldsymbol{\Psi}_{(ij)}$ satisfies, by definition $E[\boldsymbol{\Psi}_{(ij)}^{\dagger}\boldsymbol{\Phi}_{(ij)}^T] = E[\boldsymbol{\Psi}_{(ij)}^*\boldsymbol{\Phi}_{(ij)}^T]$. Hence, post-multiplying this equality by $\mathbf{C}_{ij}^T$, one gets

$E[\boldsymbol{\Psi}^{\dagger}_{(ij)}\boldsymbol{\Psi}^{T}_{(ij)}] = E[\boldsymbol{\Psi}^{*}_{(ij)}\boldsymbol{\Psi}^{T}_{(ij)}] = \mathbf{H}^{T}_{(ij)}$, yielding

$$E\left(\begin{bmatrix}\boldsymbol{\Psi}_{(ij)}\\ \boldsymbol{\Psi}^{\dagger}_{(ij)}\end{bmatrix}[\boldsymbol{\Psi}^{T}_{(ij)}\ (\boldsymbol{\Psi}^{\dagger}_{(ij)})^{T}]\right) = \begin{bmatrix}\mathbf{G}_{(ij)} & \mathbf{H}_{(ij)}\\ \mathbf{H}^{T}_{(ij)} & E[\boldsymbol{\Psi}^{\dagger}_{(ij)}(\boldsymbol{\Psi}^{\dagger}_{(ij)})^{T}]\end{bmatrix}.$$

From the nonnegativity of the matrix in the above left-hand side $\mathbf{H}_{(ij)}\mathbf{G}^{-1}_{(ij)}\mathbf{H}^{T}_{(ij)} \leq E[\boldsymbol{\Psi}^{\dagger}_{(ij)}(\boldsymbol{\Psi}^{\dagger}_{(ij)})^{T}]$ and equality is achieved if $\boldsymbol{\Psi}_{(ij)} = \boldsymbol{\Psi}^{\dagger}_{(ij)}$. Thus, one gets the result if one has proved that $\boldsymbol{\Psi}^{\dagger}_{(ij)}$ is of the form $\mathbf{C}^{*}_{(ij)}\boldsymbol{\Phi}_{(ij)}$, where $\mathbf{C}^{*}_{(ij)}$ is defined in the same way as $\mathbf{C}_{(ij)}$ with c^{*}_{in}, c^{*}_{jn} in place of c_{in}, c_{jn}. To this end, it suffices to prove the matrix equality $\mathbf{C}^{*}_{(ij)}E[\boldsymbol{\Phi}_{(ij)}\boldsymbol{\Phi}^{T}_{(ij)}] = E[\boldsymbol{\Psi}^{*}_{(ij)}\boldsymbol{\Phi}^{T}_{(ij)}]$. Now, observe that the elements of first row of the matrix in the left-hand side of this equality are $\sum_{n=1}^{N} c^{*}_{in}E[\phi_n(S_i)S_j^2\phi_m(S_i)]$, $m = 1, \cdots, N$, $\sum_{n=1}^{N} c^{*}_{in}E[\phi_n(S_i)S_j\phi_m(S_j)S_i]$, $m = 2, \cdots, N$. However, using the independence of S_i and S_j and (2.6) together with the fact that $S_i = \phi_1(S_i)$, these elements reduce to $E[\psi^{*}_i(S_i)\phi_m(S_i)]E(S_j^2) = E[\psi^{*}_i(S_i)S_j\phi_m(S_i)S_j]$, $m = 1, \cdots, N$, $E[\psi^{*}_i(S_i)S_i]E[\phi_m(S_j)S_j] = E[\psi^{*}_i(S_i)S_i\phi_m(S_j)S_j]$, $m = 2, \cdots, N$. This yields the desired matrix equality as far as the first row is concerned. The proof is similar for the other row. ■

The above result also provides a simple way to estimate the optimal choice for ψ_i or, equivalently, the coefficients c^{*}_{in}. Indeed, by (2.5), the left-hand sides of (2.6) can be replaced by $E[\phi'_m(S_i)]$. (Note that the result then becomes more general since it can be shown to continue to hold even when f^{*} is discontinuous so that ψ^{*} is undefined.) Then, replacing the expectation operator E by the time average operator $\hat{E}$, one gets a system of estimating equations for the c^{*}_{in}.

III. The ML Approach for Temporally Correlated Sources

A. Derivation of the Separation Procedure

A well-known technique to decorrelate a stationary signal is to perform a discrete Fourier transform, which transforms $\mathbf{X}(1), \cdots, \mathbf{X}(T)$ into

$$\mathbf{d}_X\left(\frac{k}{T}\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{X}(t)e^{-i2\pi tk/T}, \qquad k = 0, \cdots, T-1.$$

It is well known [1] that $E[\mathbf{d}_X(k/T)\overline{\mathbf{d}}^{T}_X(l/T)] \to 0$ as $T \to \infty$ as soon as $k \neq l$ and $E[\mathbf{d}_X(k/T)\overline{\mathbf{d}}^{T}_X(k/T)]$ converge to the same limit as the spectral density of the process at k/T. The overbar $\bar{}$ denotes the complex conjugate. Further, these random vectors are asymptotically Gaussian under mild assumptions [1]. For this reason, it is a usual practice in time series literature to regard $\{\mathbf{d}_X(k/T),\ 0 \leq k \leq T/2\}$ as actually independent Gaussian random vectors. (Note that $\mathbf{d}_X(k/T) = \overline{\mathbf{d}}_X(1 - k/T)$ since we are concerned with real signal.) This is done in particular in [18], in which a deconvolution problem having some similarities to our problem is considered. In [18], however, there is a much higher number of channels than the sources (the former are indexed by a pair i, j, whereas the latter are indexed only by a single i), and the latter are noise corrupted so that the problem is, in fact, a kind of factor analysis problem (as mentioned in Section II) *but in the frequency domain*. Note that the above practice is not quite justified here because it ignores the higher-than-second-order structure of the process. The flaw comes from the fact that the above-mentioned convergence holds only for any fixed finite subset of the frequency index k but not uniformly in $0 \leq k \leq T/2$. Nevertheless, this practice is fully justified for Gaussian signals. We must stress that it is only meant to derive a system of estimation equations, their full justification coming later, in the performance study. Thus, we take as the joint probability distribution of $\mathbf{d}_X(0)$, $\mathbf{d}_X(1/T), \cdots, \mathbf{d}_X(1/2)$, for T even as shown in (2.6a), shown at the bottom of the page, where $\mathbf{f}_X$ denotes the (matrix) spectral density of the $\mathbf{X}(t)$ process. Note that $\mathbf{d}_X(0)$ and $\mathbf{d}_X(1/2)$ are real, whereas the $\mathbf{d}_X(k/T)$, $0 < k < T/2$ are complex, Gaussian vectors. The expression for the joint density when T is odd is quite similar; the terms corresponding to $\mathbf{d}_X(1/2)$ simply disappear. Since $\mathbf{f}_X = \mathbf{A}\ \text{diag}\ (g_1, \cdots, g_K)\mathbf{A}^{T}$, where the g_i are the spectral densities of the sources and diag$(\cdot)$ denote the diagonal matrix with the indicated diagonal element, the logarithm of the above joint density reduces to

$$L_T = -\frac{1}{2}\sum_{i=1}^{K}\sum_{k=0}^{T-1}\left[\frac{\left|\mathbf{e}_i^{T}\mathbf{A}^{-1}\mathbf{d}_X\left(\frac{k}{T}\right)\right|^2}{g_i\left(\frac{k}{T}\right)} + \ln\, g_i\left(\frac{k}{T}\right)\right] - T\ln\,|\det\,\mathbf{A}| + \text{Constant}.$$

This is also the logarithm of the joint density of the data since there is a one-to-one linear map with unit Jacobian from $\mathbf{X}(1), \cdots, \mathbf{X}(T)$ to $\mathbf{d}_X(k/T)$, $0 \leq k \leq T/2$.

We make (for the moment) the working assumption that the spectral densities of the sources are known up to a constant. Thus, $g_i(k/T) = \lambda_i h_i(k/T)$, where h_i are known functions, and λ_i are unknown parameters. The ML approach consists

$$\frac{1}{\sqrt{\det\,\left[4\pi^2\mathbf{f}_X(0)\mathbf{f}_X(\tfrac{1}{2})\right]^{1/2}}}\ \exp\left[-\frac{\mathbf{d}_X^{T}(0)\mathbf{f}_X^{-1}(0)\mathbf{d}_X(0) + \mathbf{d}_X^{T}(\tfrac{1}{2})\mathbf{f}_X^{-1}(\tfrac{1}{2})\mathbf{d}_X(0)}{2}\right] \cdot \prod_{k=1}^{T/2-1}\frac{1}{\det\left[\pi\mathbf{f}_X\left(\frac{k}{T}\right)\right]}\ \exp\left[-\overline{\mathbf{d}}_X^{T}\left(\frac{k}{T}\right)\mathbf{f}_X^{-1}\left(\frac{k}{T}\right)\mathbf{d}_X\left(\frac{k}{T}\right)\right] \tag{2.6a}$$

in maximizing L_T with respect to $\lambda_1, \cdots, \lambda_K$ and $\mathbf{A}$. As in Section II, we shall consider the differential of this function. From $d\mathbf{A}^{-1} = -\mathbf{A}^{-1}\,d\mathbf{A}\mathbf{A}^{-1}$, $d \ln |\det \mathbf{A}| = tr(\mathbf{A}^{-1}\,d\mathbf{A})$, we get

$$\begin{aligned} dL_T &= \sum_{i=1}^{K}\sum_{k=0}^{T-1} \frac{\mathbf{e}_i^T\mathbf{A}^{-1}\,d\mathbf{A}\mathbf{A}^{-1}\,\mathbf{d}_X\left(\frac{k}{T}\right)\mathbf{e}_i^T\mathbf{A}^{-1}\,\overline{\mathbf{d}}_X\left(\frac{k}{T}\right)}{\lambda_i h_i\left(\frac{k}{T}\right)} \\ &\quad - T\,tr(\mathbf{A}^{-1}\,d\mathbf{A}) \\ &\quad + \frac{1}{2}\sum_{i=1}^{K}\left[\sum_{k=0}^{T-1}\frac{\left|\mathbf{e}_i^T\mathbf{A}^{-1}\,\mathbf{d}_X\left(\frac{k}{T}\right)\right|^2}{\lambda_i h_i\left(\frac{k}{T}\right)} - T\right]\frac{d\lambda_i}{\lambda_i}. \end{aligned}$$

Introducing the general element ∂_{ij} of the matrix $\mathbf{A}^{-1}\,d\mathbf{A}$, the first two terms of the above right-hand side can be rewritten as

$$\sum_{i=1}^{K}\sum_{j=1}^{K}\left[\sum_{k=0}^{T-1}\frac{\mathbf{e}_j^T\mathbf{A}^{-1}\,\mathbf{d}_X\left(\frac{k}{T}\right)\mathbf{e}_i^T\mathbf{A}^{-1}\,\overline{\mathbf{d}}_X\left(\frac{k}{T}\right)}{\lambda_i h_i\left(\frac{k}{T}\right)}\right]\partial_{ij} - T\sum_{i=1}^{k}\partial_{ii}.$$

At the ML estimators $\hat{\mathbf{A}}$, $\hat{\lambda}_1, \cdots, \hat{\lambda}_K$, the differential of L_T must vanish identically for all infinitesimal increments $d\mathbf{A}$, $d\lambda_1, \cdots, d\lambda_K$ of its arguments. This yields the estimating equations

$$\sum_{k=0}^{T-1}\frac{\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\,\mathbf{d}_X\left(\frac{k}{T}\right)\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\,\overline{\mathbf{d}}_X\left(\frac{k}{T}\right)}{h_i\left(\frac{k}{T}\right)} = 0 \qquad i \neq j = 1, \cdots, K$$

$$\hat{\lambda}_i = \frac{1}{T}\sum_{k=0}^{T-1}\frac{\left|\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\,\mathbf{d}_X\left(\frac{k}{T}\right)\right|^2}{h_i\left(\frac{k}{T}\right)}, \qquad i = 1, \cdots, K.$$

The second set of the above equations merely serves to determine the scale factors and is of no interest. The estimated mixing matrix can be determined (up to a scale factor) by the first set.

It is, however, simpler to revert to the time domain, slightly modifying the estimating equations. Approximating sums by integrals, they can be written approximately as

$$\int_0^1 \mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\,\mathbf{d}_X(\nu)\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\,\overline{\mathbf{d}}_X(\nu)h_i^{-1}(\nu)\,d\nu = 0, \qquad i \neq j = 1, \cdots, K.$$

Note that $\sqrt{T}\,\mathbf{d}_X(\nu)/h_i(\nu) = [\sum_{t=1}^{T}\mathbf{X}(t)e^{-i2\pi t\nu}]/h_i(\nu)$ is the Fourier series with coefficients the convolution of $\mathbf{X}(t)$ with the linear filter ϕ_i with frequency response h_i^{-1} (by the convention $\mathbf{X}(t) = 0$ for $t < 1$ or $t > T$). Thus, by the Parseval equality, the above equations can be written as follows, where $*$ denotes the convolution operator:

$$\hat{E}[\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\phi_i * \mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X})] = 0, \qquad i \neq j = 1, \cdots, K \tag{3.1}$$

and where $\hat{E}$ is the time average operator as in Section II-A. Note that the filter ϕ_i may have an infinite impulse response so that $\phi_i * \mathbf{X}$ will involve the $\mathbf{X}(t)$ for all t, but those $\mathbf{X}(t)$ for $t \notin \{1, \cdots, T\}$ have been put to zero. In the case where the ϕ_i have a finite impulse response (FIR), one may choose to restrict the averaging (3.1) to a smaller range so that the $(\phi_i * \mathbf{X})(t)$ for t in this range might be computed from $\mathbf{X}(1), \cdots, \mathbf{X}(T)$.

The interpretation of (3.1) is clear: The estimator $\hat{\mathbf{A}}$ is such that each reconstructed source $\hat{S}_i(t) = \mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X}(t)$, transformed by the linear filter ϕ_i, is empirically uncorrelated in the sense of the time average operator $\hat{E}$, with all other reconstructed sources.

In practice, the spectral densities of the sources are unknown so that we cannot take ϕ_i as above but have to choose them *a priori*. Such a procedure is still justified since, as is seen from the above interpretation, it only exploits the fact that any filtered version of a source is uncorrelated with other sources and that the time average of a product of such two processes converges to its expectation as the time interval increases to infinity. Note that there are $K(K-1)$ estimating equations in (3.1), which is just enough to estimate the matrix $\mathbf{A}$, up to scaling. Of course, the choice of the filters ϕ_i will affect the performance of the method. In the next section, we will evaluate the asymptotic performance and show that it is highest when these filters are chosen to have their frequency responses proportional to the inverses of the spectral densities of the sources.

We now derive some equivalent forms of (3.1). Let $\phi_{i,k}$, $k = \cdots, -1, 0, 1, \cdots$ denote the impulse response of the filter ϕ_i; then, noting that $\mathbf{X}(t)$ has been put to zero for $t \notin \{1, \cdots, T\}$, (3.1) can be written explicitly as

$$\frac{1}{T}\sum_{k=1-T}^{T-1}\phi_{i,k}\sum_{t=\max(1,1-k)}^{\min(T,T-k)}\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}(t)\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X}(t-k) = 0, \qquad i \neq j = 1, \cdots, K.$$

Thus, letting

$$\mathbf{R}_T(k) = \frac{1}{T}\sum_{t=\max(1,1-k)}^{\min(T,T-k)}\mathbf{X}(t)\mathbf{X}^T(t-k) \qquad [\mathbf{R}_T(k) = 0 \text{ for } |k| \geq T]$$

denote the sample autocovariance matrices of the observed process, the estimating equations for $\mathbf{A}$ become

$$\sum_{k=1-T}^{T-1}\phi_{i,k}\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{R}_T(k)\hat{\mathbf{A}}^{-T}\mathbf{e}_i = 0, \qquad i \neq j = 1, \ldots, K.$$

The above system of equations is particularly interesting when ϕ_i are FIR filters: $\phi_{i,k} = 0$ for $|k| > p$, say, in which

case, all computations can be done through the matrices $\mathbf{R}_T(0), \cdots, \mathbf{R}_T(p)$.

Let us restrict ϕ_i to the form $\tilde{\phi}_i * \check{\tilde{\phi}}_i$, where $\tilde{\phi}_i$ is causal, and $\check{\tilde{\phi}}_i$ is its mirror obtained by reversing the time order of its impulse response. Such a restriction is justified by the fact that the optimal choice for ϕ_i is the filter with frequency response inversely proportional to the spectral density of the ith source and, thus, can be written in the above form, provided that the logarithm of its frequency response in integrable (since, then, this frequency response, being even and nonnegative, can be written as the squared modulus of a one-sided Fourier series). Now, it can be checked that for any two sequences $y(t)$ and $z(t)$, $\sum_{t=-\infty}^{\infty} y(t)(\check{\tilde{\phi}}_i * z)(t) = \sum_{t=-\infty}^{\infty}(\tilde{\phi}_i * y)(t)z(t)$. Thus, an equivalent form of (3.1) is

$$\hat{E}[(\tilde{\phi}_i * \mathbf{e}_j^T \hat{\mathbf{A}}^{-1}\mathbf{X})(\tilde{\phi}_i * \mathbf{e}_i^T \hat{\mathbf{A}}^{-1}\mathbf{X})] = 0, \qquad i \neq j = 1, \cdots, K.$$

When the filter ϕ_i has frequency response function proportional to the inverse of the spectral density of the ith source, $\tilde{\phi}_i$ is, in fact, the whitening filter for this source (but not for other sources). In this case, our estimation procedure can be interpreted as making *each reconstructed source whitened through a whitening filter be empirically uncorrelated with the filtered versions of all other sources by the same filter*. However, in practice, the spectral densities of the sources are unknown, and the above filters are only a guess (or some estimate) of the true whitening filters and need not really whiten the sources.

B. Asymptotic Properties of the Estimator

We shall use the same approach as in Section II-B to study the asymptotic properties of the estimator. Thus, we write $\hat{\mathbf{A}} = \mathbf{A}(\mathbf{I} - \boldsymbol{\delta})^{-1}$, where $\boldsymbol{\delta}$ is a small matrix. Then, $\mathbf{e}_i^T \hat{\mathbf{A}}^{-1}\mathbf{X}(t) = S_i(t) - \sum_{k=1}^{K} \delta_{ik} S_k(t)$, where $S_i(t) = \mathbf{A}^{-1}\mathbf{X}(t)$, and δ_{ij} denotes the general element of the matrix $\boldsymbol{\delta}$. Therefore, the right hand side of (3.1) can be written approximately as

$$\hat{E}[S_j(\phi_i * S_i)] - \sum_{k=1}^{K} \{\hat{E}[S_k(\phi_i * S_i)]\delta_{jk} + \hat{E}[S_j(\phi_i * S_k]\delta_{ik}\}.$$

Assuming ergodicity, $\hat{E}[S_j\ (\phi_i * S_k)]$ converges as $T \to \infty$ to $E[S_j(t)(\phi_i * S_k)(t)]$, which vanishes if $j \neq k$ and equals $\int_{-1/2}^{1/2} g_j^*(\nu) h_i^{-1}(\nu)\, d\nu$ otherwise, where g_j^* denotes the true spectral density of the signal $S_j(t)$, and $1/h_i$ denottes the frequency response of the filter ϕ_i. Thus, the system of (3.1) can be rewritten approximately as

$$\left(\int \frac{g_i^*}{h_i} d\nu\right)\delta_{ji} + \left(\int \frac{g_j^*}{h_i} d\nu\right)\delta_{ij} \approx \hat{E}[S_j(\phi_i * S_i)], \qquad i \neq j = 1, \cdots, K.$$

For simplicity, the integration variable ν and its range have been and will be dropped. To avoid new notations, we shall reuse some of Section II-B for different but similar quantities. We put $\boldsymbol{\Delta}$ and $\boldsymbol{\Psi}$ as the $K(K-1)$-vector $(\delta_{12}\ \delta_{21}\ \cdots)^T$ and $(\Psi_{12}\ \Psi_{21}\ \cdots)^T$, where $\Psi_{ij} = S_j(\phi_i * S_i)$. Then, the above system of equations becomes $\boldsymbol{\Delta} \approx \mathbf{H}^{-1}\hat{E}(\boldsymbol{\Psi})$, where $\mathbf{H}$ is the bloc diagonal matrix with blocks

$$\mathbf{H}_{(ij)} = \begin{bmatrix} \int \frac{g_j^*}{h_i} d\nu & \int \frac{g_i^*}{h_i} d\nu \\ \int \frac{g_j^*}{h_j} d\nu & \int \frac{g_i^*}{h_j} d\nu \end{bmatrix}$$

indexed by the pairs (i, j), $i \neq j = 1, \cdots, K$. On the other hand, under mild conditions, the central limit theorem can be applied to the average of the stationary process $\boldsymbol{\Psi}(t)$ to yield that $\sqrt{T}\hat{E}(\boldsymbol{\Psi})$ converges in distribution as $T \to \infty$ to a Gaussian random vector with zero mean and covariance matrix [16]

$$\mathbf{G} = \lim_{T\to\infty} \frac{1}{T} E\left[\sum_{t=1}^{T}\sum_{s=1}^{T} \boldsymbol{\Psi}(t)\boldsymbol{\Psi}(s)^T(t)\right].$$

To compute $\mathbf{G}$, we need the autocovariance functions $\gamma_i(k) = E[S_i(t)S_i(t+k)]$ and $\tilde{\gamma}_i(k) = E[(\phi_i * S_i)(t)(\phi_i * S_i)(t+k)]$ of the processes $S_i(t)$ and $(\phi_i * S_i)(t)$ and the cross-covariance function $\tilde{\gamma}_i(k) = E[S_i(t)(\phi_i * S_i)(t+k)]$ between these processes. Using the mutual independence between the zero mean stationary processes $S_i(t)$, the matrix $\mathbf{G}$ can be seen to be block diagonal with blocks

$$\mathbf{G}_{(ij)} = \sum_{k=-\infty}^{\infty} \begin{bmatrix} \gamma_j(k)\tilde{\gamma}_i(k) & \tilde{\gamma}_j(k)\tilde{\gamma}_i(-k) \\ \tilde{\gamma}_i(k)\tilde{\gamma}_j(-k) & \gamma_i(k)\tilde{\gamma}_j(k) \end{bmatrix}$$

indexed by the pairs (i, j), $i \neq j = 1, \cdots, K$. Since the spectral density of the process $(\phi_i * S_i)(t)$ is g_i^*/h_i^2 and the cross-spectral density between $S_i(t)$ and $(\phi_i * S_i)(t)$ is g_i^*/h_i, by the Parseval equality

$$\mathbf{G}_{(ij)} = \begin{bmatrix} \int \frac{g_j^* g_i^*}{h_i^2\, dv} & \int \frac{g_j^* g_i^*}{h_j h_i\, dv} \\ \int \frac{g_i^* g_j^*}{h_i h_j\, dv} & \int \frac{g_i^* g_j^*}{h_j^2\, dv} \end{bmatrix}, \qquad i \neq j.$$

Finally, since both $\mathbf{H}$ and $\mathbf{G}$ are block diagonal, the $(\delta_{ij}\ \delta_{ji})^T$ for different pairs (i, j) are asymptotically independent with covariance matrix $\mathbf{H}_{(ij)}^{-1}\mathbf{G}_{(ij)}(\mathbf{H}_{(ij)}^{-1})^T/T$.

We now show that the above covariance matrix is minimum when the filters ϕ_i are chosen so that their frequency response h_i^{-1} are inversely proportional to the true spectral densities g_i^* of the sources. Observe that the matrices $\mathbf{H}_{(ij)}$ and $\mathbf{G}_{(ij)}$ can be written as

$$\int \begin{bmatrix} \frac{1}{h_i} \\ \frac{1}{h_j} \end{bmatrix} \begin{bmatrix} \frac{1}{g_i^*} & \frac{1}{g_j^*} \end{bmatrix} g_i^* g_j^*\, d\nu$$

and

$$\int \begin{bmatrix} \frac{1}{h_i} \\ \frac{1}{h_j} \end{bmatrix} \begin{bmatrix} \frac{1}{h_i} & \frac{1}{h_j} \end{bmatrix} g_i^* g_j^*\, d\nu.$$

Thus, defining

$$\mathbf{J}_{(ij)} = \int \begin{bmatrix} \frac{1}{g_i^*} \\ \frac{1}{g_j^*} \end{bmatrix} \begin{bmatrix} \frac{1}{g_i^*} & \frac{1}{g_j^*} \end{bmatrix} g_i^* g_j^* \, d\nu = \begin{bmatrix} \int \frac{g_j^*}{g_i} d\nu & 1 \\ 1 & \int \frac{g_i^*}{g_j^*} d\nu \end{bmatrix}$$

one has

$$\int \begin{bmatrix} \frac{1}{h_i} \\ \frac{1}{h_j} \\ \frac{1}{g_i^*} \\ \frac{1}{g_j^*} \end{bmatrix} \begin{bmatrix} \frac{1}{h_i} & \frac{1}{h_j} & \frac{1}{g_i^*} & \frac{1}{g_j^*} \end{bmatrix} g_i^* g_j^* d\nu = \begin{bmatrix} \mathbf{G}_{(ij)} & \mathbf{H}_{(ij)} \\ \mathbf{H}_{(ij)}^T & \mathbf{J}_{(ij)} \end{bmatrix}.$$

From the nonnegativity of the matrix in the left-hand side of the above equality, $\mathbf{H}_{(ij)}^{-1}\mathbf{G}_{(ij)}(\mathbf{H}_{(ij)}^{-1})^T \geq \mathbf{J}_{(ij)}^{-1}$. Equality can be attained if each of the h_i is proportional to the corresponding g_i^*. This yields the announced result. Further, the minimum asymptotic covariance matrix of the estimator is given by $\mathbf{J}^{-1}/T$, where $\mathbf{J}$ is also a block diagonal matrix with blocks $\mathbf{J}_{(ij)}$ indexed by the pairs (i, j), $i \neq j = 1, \cdots, K$.

The above results are derived under the implicit assumption that the matrix $\mathbf{H}_{(ij)}$ is invertible. In order that it be so, it is necessary (but not sufficient) that h_i and h_j, on the one hand, and g_i^* and g_j^*, on the other hand, are not proportional since otherwise, $\mathbf{H}_{(ij)}$ would have proportional rows or proportional columns. Interestingly, the asymptotic covariance matrix of $(\delta_{ij} \;\; \delta_{ji})^T$ depends only on the linear space spanned by h_i^{-1} and h_j^{-1} and not on these functions themselves; hence, they could be chosen nearly proportional without affecting the asymptotic performance of the method (except that some numerical instability may arise). To see this, suppose that another set of filters have been chosen with frequency responses as the components of $\mathbf{C}_{(ij)}(h_i^{-1} \;\; h_j^{-1})^T$, where $\mathbf{C}_{(ij)}$ is a nonsingular matrix (since otherwise these responses would be proportional). Then, a simple computation shows that the new matrices $\mathbf{H}_{(ij)}$ and $\mathbf{G}_{(ij)}$ associated with this set of filters are given by $\tilde{\mathbf{H}}_{(ij)} = \mathbf{C}_{(ij)}\mathbf{H}_{(ij)}$, $\tilde{\mathbf{G}}_{(ij)} = \mathbf{C}_{(ij)}\mathbf{G}_{(ij)}\mathbf{C}_{(ij)}^T$; hence, the asymptotic covariance matrix of $(\delta_{ij} \;\; \delta_{ji})^T$ remains the same. On the other hand, it is essential that the true spectral densities g_i^* and g_j^* be not too similar for the procedure to perform well. Indeed, the determinant of the matrix $\mathbf{J}_{(ij)}$, from the above result, can be shown to be equal to

$$\int \frac{g_j^*}{g_i^*} d\nu \int \frac{g_i^*}{g_j^*} d\nu - 1 = \int \frac{\left[\frac{g_i^*}{g_j^*} - \int \frac{g_i^*}{g_j^*} d\nu\right]^2}{\frac{g_i^*}{g_j^*} \int \frac{g_i^*}{g_j^*} d\nu} d\nu.$$

Hence, the diagonal elements of the matrix $\mathbf{J}_{(ij)}^{-1}$ are

$$\left[\int \frac{\left(\frac{g_i^*}{g_j^*} - \int \frac{g_i^*}{g_j^*} d\nu\right)^2}{\frac{g_i^*}{g_j^*}\left(\int \frac{g_i^*}{g_j^*} d\nu\right)^2} d\nu\right]^{-1}$$

and

$$\left[\int \frac{\left(\frac{g_j^*}{g_i^*} - \int \frac{g_j^*}{g_i^*} d\nu\right)^2}{\frac{g_j^*}{g_i^*}\left(\int \frac{g_j^*}{g_i^*} d\nu\right)^2} d\nu\right]^{-1}$$

which will be large if g_i^*/g_j^* is almost constant, i.e., close to its integral on $[-1/2, 1/2]$.

C. Construction of the Estimator

1) General Iterative Method: The same approach as in Section II-C can be used to derive an iterative algorithm for finding the solution of (3.1). By expanding (3.1) around an initial estimate $\tilde{\mathbf{A}}$ of $\mathbf{A}$ and using the same computation as in Section III-B, the next step estimate, which is denoted by $\tilde{\mathbf{A}}(\mathbf{I} - \tilde{\boldsymbol{\delta}})^{-1}$, can be constructed as the solution of

$$\sum_{k=1}^{K} \{\hat{E}[\tilde{S}_j(\phi_i * \tilde{S}_k)]\tilde{\delta}_{ik} + \hat{E}[\tilde{S}_k(\phi_i * \tilde{S}_i)]\tilde{\delta}_{jk}\} = \hat{E}[\tilde{S}_j(\phi_i * \tilde{S}_i)]$$

where $\tilde{S}_i = \mathbf{e}_i^T \tilde{\mathbf{A}}^{-1}\mathbf{X}$, and $\tilde{\delta}_{ij}$ denotes the general element of $\tilde{\boldsymbol{\delta}}$. Since $\tilde{S}_i(t)$ is close to $S_i(t)$, one may approximate $\hat{E}[\tilde{S}_j(\phi_i * \tilde{S}_k)]$ by $E[S_j(t)(\phi_i * S_k)(t)]$, which vanishes if $k \neq j$. Thus, the $\tilde{\delta}_{ij}$ can be pairwise computed through the following system of equations:

$$\hat{E}[\tilde{S}_j(\phi_i * \tilde{S}_j)]\tilde{\delta}_{ij} + \hat{E}[\tilde{S}_i(\phi_i * \tilde{S}_i)]\tilde{\delta}_{ji} = \hat{E}[\tilde{S}_j(\phi_i * \tilde{S}_i)], \qquad i \neq j = 1, \cdots, K. \tag{3.2}$$

This yields the new estimator of $\mathbf{A}$, and one can iterate the procedure until convergence. As before, when the ϕ_i are FIR filters of order p, all computations can be done through the sample covariance matrices $\mathbf{R}_T(0), \cdots, \mathbf{R}_T(p)$. Indeed

$$\hat{E}[\tilde{S}_l(\phi_i * \tilde{S}_m)] = \mathbf{e}_l^T \tilde{\mathbf{A}}^{-1} \left[\sum_{k=-p}^{p} \phi_{i,k}\mathbf{R}_T(k)\right] \tilde{\mathbf{A}}^{-T}\mathbf{e}_m \qquad l, m = 1, \cdots, K \tag{3.3}$$

with $\mathbf{R}_T(-k) = \mathbf{R}_T^T(k)$. In addition, when the filter ϕ_i is of the form $\check{\tilde{\phi}}_i * \tilde{\phi}_i$ with $\tilde{\phi}_i$ a causal filter and $\check{\tilde{\phi}}_i$ its mirror image:

$$\hat{E}[\tilde{S}_l(\phi_i * \tilde{S}_m)] = \hat{E}[(\tilde{\phi}_i * \tilde{S}_l)(\tilde{\phi}_i * \tilde{S}_m)] \qquad l, m = 1, \cdots, K. \tag{3.4}$$

We now discuss the modification of the above procedure for use in an adaptive context. Let $\tilde{\mathbf{A}}(t-1)$ be an estimate of $\mathbf{A}$ at time $t-1$. As before, the new estimate at time t is $\tilde{\mathbf{A}}(t) = \tilde{\mathbf{A}}(t-1)[\mathbf{I} - \tilde{\boldsymbol{\delta}}(t)]^{-1}$, where $\tilde{\boldsymbol{\delta}}(t)$ is the matrix with

general element $\tilde{\delta}_{ij}(t)$, which is the solution of an equation analogous to (3.2):

$$\Omega_{ij}(t)\tilde{\delta}_{ij}(t)+\Omega_{ii}(t)\tilde{\delta}_{ji}(t)=\omega_{ij}(t). \qquad i\neq j=1,\cdots,K.$$

Concerning the definition of $\Omega_{ij}(t)$ and $\omega_{ij}(t)$, there are two possibilities. First, in the case where the filters ϕ_i are FIR, using (3.2) and (3.3), one may define

$$\Omega_{ij}(t)=\mathbf{e}_j^T\tilde{\mathbf{A}}^{-1}(t-1)\left[\sum_{k=-p}^{p}\phi_{i,k}\mathbf{P}(k,t)\right]\cdot\tilde{\mathbf{A}}^{-1}(t-1)^T\mathbf{e}_j, \qquad i,j=1,\cdots,K.$$

$$\omega_{ij}(t)=\mathbf{e}_j^T\tilde{\mathbf{A}}^{-1}(t-1)\left[\sum_{k=-p}^{p}\phi_{i,k}\mathbf{P}(k,t)\right]\cdot\tilde{\mathbf{A}}^{-1}(t-1)^T\mathbf{e}_i, \qquad i\neq j=1,\cdots,K$$

where $\mathbf{P}(k,t)$ is the adaptive estimate of the autocovariance matrix of lag k up to the factor $1/(1-\lambda)$, which is defined by the recursion

$$\mathbf{P}(k,t)=\lambda\mathbf{P}(k,t-1)+\mathbf{X}(t)\mathbf{X}^T(t-k)$$

where λ is a forgetting factor. In the case where ϕ_i is the convolution between a causal filter $\tilde{\phi}_i$ and its mirror image, an equivalent procedure consists of directly updating the matrices $\mathbf{P}^{(i)}(t)=\sum_{k=-p}^{p}\phi_{i,k}\mathbf{P}(k,t)$, as follows:

$$\mathbf{P}^{(i)}(t)=\mathbf{P}^{(i)}(t-1)+(\tilde{\phi}_i*\mathbf{X})(t)(\tilde{\phi}_i*\mathbf{X}^T)(t).$$

The second possibility consists of updating directly the $\Omega_{ij}(t)$ and $\omega_{ij}(t)$, based on (3.2) and (3.4):

$$\Omega_{ij}(t)=\lambda\Omega_{ij}(t-1)+[(\tilde{\phi}_i*\tilde{S}_j)(t)]^2 \qquad i,j=1,\cdots,K$$

$$\omega_{ij}(t)=(\tilde{\phi}_i*\tilde{S}_j)(t)(\tilde{\phi}_i*\tilde{S}_i)(t) \qquad i\neq j=1,\cdots,K$$

where $\tilde{S}_i(t)=\mathbf{e}_i^T\tilde{\mathbf{A}}(t-1)\mathbf{X}(t)$. This approach is quite similar to the adaptive algorithm of Section II-C.

2) Direct Computation: There is, in particular, an interesting case where our estimator *can be computed directly without any iteration*. This happens when the filters ϕ_i are symmetric and belong to a linear space of dimension two. More precisely, suppose that $\phi_i=\alpha_{i1}\varphi_1+\alpha_{i2}\varphi_2$ for two symmetric filters φ_1 and φ_2 and some coefficients α_{i1}, α_{i2} $(i=1,\cdots,k)$. Then, the estimating equations (3.1) can be written as

$$\alpha_{i,1}\hat{E}[\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\varphi_1*\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X})+\alpha_{i,2}\hat{E}[\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\varphi_2*\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X})=0,$$

$$\alpha_{j,1}\hat{E}[\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\varphi_1*\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X})+\alpha_{j,2}\hat{E}[\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\varphi_2*\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X})=0$$

where we have grouped the indices into pairs (i,j), (j,i), $1\leq i<j\leq K$. However, $\hat{E}[\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\varphi_m*\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X})]=\hat{E}[\mathbf{e}_i^T(\varphi_m*\hat{\mathbf{A}}^{-1}\mathbf{X})\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}]$, $m=1,2$ because φ_1 and φ_2 are symmetric; hence, the above pair of equations takes the form

$$\begin{bmatrix}\alpha_{i1} & \alpha_{i2}\\ \alpha_{j1} & \alpha_{j2}\end{bmatrix}\begin{bmatrix}\hat{E}[\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\varphi_1*\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X})]\\ \hat{E}[\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\varphi_2*\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X})]\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}.$$

Note that for all $i\neq j$, the row vectors $(\alpha_{i1}\ \alpha_{i2})$ and $(\alpha_{j1}\ \alpha_{j2})$ are not proportional since the filters ϕ_i and ϕ_j must not have proportional frequency responses. Thus, the above matrix equation is equivalent to

$$\begin{bmatrix}\hat{E}[\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\varphi_1*\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X})]\\ \hat{E}[\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\mathbf{X}(\varphi_2*\mathbf{e}_i^T\hat{\mathbf{A}}^{-1}\mathbf{X})]\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}.$$

Since the linear filtering and the operator $\hat{E}$ commute with the linear instantaneous transformation, one finally gets that (3.1) is equivalent to

$$\mathbf{e}_j^T\hat{\mathbf{A}}^{-1}\hat{E}[\mathbf{X}(\varphi_m*\mathbf{X})^T]\hat{\mathbf{A}}^{-T}\mathbf{e}_i=0, \qquad i\neq j=1,\cdots,K, m=1,2. \tag{3.5}$$

Introducing the matrices $\mathbf{Q}_T^{(m)}=\hat{E}[\mathbf{X}(\varphi_m*\mathbf{X})^T]=\hat{E}[(\varphi_m*\mathbf{X})\mathbf{X}^T]$, $m=1,2$, (3.5) indicates that the matrices $\hat{\mathbf{A}}^{-1}\mathbf{Q}_T^{(1)}\hat{\mathbf{A}}^{-T}$ and $\hat{\mathbf{A}}^{-1}\mathbf{Q}_T^{(2)}\hat{\mathbf{A}}^{-T}$ are diagonal. Thus, *the estimator* $\hat{\mathbf{A}}$ *can be defined by the condition that its inverse simultaneously diagonalizes the matrices* $\mathbf{Q}_T^{(1)}$ and $\mathbf{Q}_T^{(2)}$. This idea of joint diagonalization actually has been used in [19]. The problem of joint diagonalization of two symmetric matrices $\mathbf{P}$, $\mathbf{Q}$ is well known. It has a solution if $\mathbf{P}$ or $\mathbf{Q}$ is positive definite [8, p. 467]. A solution, when $\mathbf{Q}$ is positive definite, is $\mathbf{UL}^{-1}$, where $\mathbf{Q}=\mathbf{LL}^T$ is any factorization $\mathbf{Q}$, and $\mathbf{U}$ is the orthogonal matrix diagonalizing $\mathbf{L}^{-1}\mathbf{PL}^{-T}$. Further, if the eigenvalues of the last matrix, which are the zeros of $\det(\mathbf{P}-\lambda\mathbf{Q})$, are distinct, then this solution is *unique* up to a permutation and a scaling. Indeed, by assumption, $\mathbf{P}=\mathbf{LU}^T\mathbf{\Lambda}\mathbf{UL}^T$, where $\mathbf{\Lambda}$ is the diagonal matrix with distinct diagonal elements. Now, for any other solution $\mathbf{T}$, the matrix $\mathbf{V}=\mathbf{TLU}^T$ satisfies the fact that $\mathbf{VV}^T$ and $\mathbf{V\Lambda V}^T$ are diagonal. Thus, the row vectors $\mathbf{v}_i$ of $\mathbf{V}$ are orthogonal, and for each i, $\mathbf{v}_i\mathbf{\Lambda}$ is also orthogonal to all other $\mathbf{v}_j$. Hence, $\mathbf{v}_i\mathbf{\Lambda}$ must be proportional to $\mathbf{v}_i$, but the diagonal elements of $\mathbf{\Lambda}$ being distinct, this is possible only if $\mathbf{v}_i$ has exactly one nonzero component, meaning that it is a permuted diagonal matrix.

Thus, our method generally provides a unique and closed-form estimator if $\mathbf{Q}_T^{(1)}$ or $\mathbf{Q}_T^{(2)}$ is positive definite. The positive definiteness of $\mathbf{Q}_T^{(m)}$ is guaranteed if one chooses the filter φ_m to have positive frequency response (which is a natural choice since the inverse of a spectral density is positive). This is because φ_m can be written as the convolution of some filter $\tilde{\varphi}_m$ with its mirror image $\check{\tilde{\varphi}}_m$ so that by the same computation as at the end of Section III-A, $\mathbf{Q}_T^{(m)}=\hat{E}[\tilde{\varphi}_m*\mathbf{X})(\tilde{\varphi}_m*\mathbf{X})^T]$ is the sample covariance matrix of the filtered process $(\tilde{\varphi}_m*\mathbf{X})(t)$. Note that when φ_m is a FIR filter, $\mathbf{Q}_T^{(m)}$ is actually a linear combination of sample covariance matrix. The idea to consider such a linear combination to get a positive definite matrix has been used in [20].

The simplest situation where the filters $\phi_1,\cdots,\phi_K$ belong to a linear space of dimension 2 is when they all are symmetric

FIR filters of first order. Then, one can takes, for example, φ_1 and φ_2 to be the filter with impulse response $\varphi_{1,k} = 2$ for $k = 0, = 1$ for $k = \pm 1, = 0$ otherwise, and $\varphi_{2,k} = 2$, for $k = 0, = -1$ for $k = \pm 1, = 0$ otherwise. Our estimation procedure then consists simply in finding a matrix $\hat{\mathbf{A}}$ such that its inverse diagonalizes simultaneously $2\mathbf{R}_T(0) + \mathbf{R}_T(1) + \mathbf{R}_T(-1)$ and $2\mathbf{R}_T(0) - \mathbf{R}_T(1) - \mathbf{R}_T(-1)$, which are the sample covariance matrices of the processes $\mathbf{X}(t) + \mathbf{X}(t-1)$ and $\mathbf{X}(t) - \mathbf{X}(t-1)$, respectively. Since the frequency response of φ_1 and φ_2 span the space of trigonometric polynomials of first degree, this simple procedure is optimal in the case where the spectral densities of the sources are first-order autoregressive processes.

Another simple situation where our joint diagonalization applies is when there are only two sources ($k = 2$). Then, clearly, one can take $\varphi_1 = \phi_1$ and $\varphi_2 = \phi_2$, and one is led to the problem of diagonalizing simultaneously two 2×2 matrices. In this simple case, the solution can be written down explicitly. Let

$$\mathbf{P} = \begin{bmatrix} p_1 & p \\ p & p_2 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} q_1 & q \\ q & q_2 \end{bmatrix}$$

be two symmetric matrices of order two. Then, it can be checked that in the matrix $\mathbf{T}$ in (3.6), shown at the bottom of the page, sign (x) denotes the sign of x, and

$$\Delta = (p_2 q_1 - p_1 q_2)^2 + 4(pq_2 - p_2 q)(pq_1 - p_1 q)$$

diagonalizes simultaneously $\mathbf{P}$ and $\mathbf{Q}$. The only required condition is that $\Delta > 0$ (if $\Delta = 0$, $\mathbf{T}$ still diagonalizes $\mathbf{P}$ and $\mathbf{Q}$, but it is singular). If $\mathbf{P}$ and $\mathbf{Q}$ are not proportional, this condition can be shown to be satisfied if at least one of these matrices has a positive determinant (which is weaker than the condition that at least one of them is positive definite). It can also be proved that the above solution is unique, up to a premultiplication by a permuted diagonal matrix. In fact, we have used this indetermination to choose $\mathbf{T}$ such that its diagonal elements are equal and greater than the square root of the absolute value of the product of its off-diagonal elements (so that the ith source ($i = 1, 2$) is the one "mostly contributed" to the ith sensor).

From the computational point of view, it is of interest to choose φ_1, φ_2 to be FIR filters of order p, say. Then, the matrices $\mathbf{Q}_T^{(1)}$ and $\mathbf{Q}_T^{(2)}$ can again be computed from the sample covariance matrices $\mathbf{R}_T(0), \cdots, \mathbf{R}_T(p)$.

The above direct method can be used in an adaptive context by updating the matrices $\mathbf{Q}_T^{(1)}$ and $\mathbf{Q}_T^{(2)}$ and then diagonalizing them. If φ_1, φ_2 are FIR filters, this can be done by updating the sample covariance matrices and if these filters can be factored as the convolution of a causal filter $\tilde{\varphi}_m$ with its mirror image ($m = 1, 2$), $\mathbf{Q}_T^{(1)}$, and $\mathbf{Q}_T^{(2)}$ can also be updated directly using the fact that $\mathbf{Q}_T^{(m)} = \hat{E}[\tilde{\varphi}_m * \mathbf{X})(\tilde{\varphi}_m * \mathbf{X})^T]$ is the sample covariance matrix of the filtered process $(\tilde{\varphi}_m * \mathbf{X})(t)$

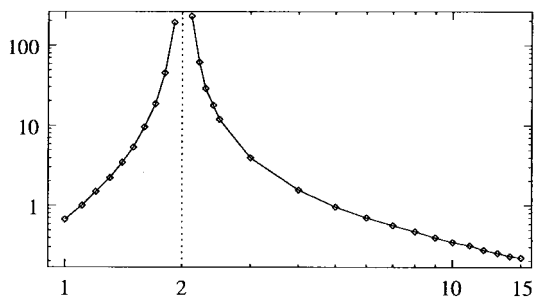

Fig. 1. Minimum asymptotic variance of the estimator.

IV. Some Numerical and Simulation Results

We give here some numerical calculations concerning the choice of the separating functions and some simulation examples. The simulations here are intended only as illustrations of the method but by no means constitute an empirical performance study (which could be the topic of a separate work).

A. Choice of Separating Functions

To study the effect of choice of separating functions, we consider the case where the sources have a same distribution taken from the family of densities $f_p(s) = K_p \exp(-|s|^p/p)$, $p \geq 1$ being a shape parameter and K_p being the normalization constant. The reason for considering this family is that it makes the analytical computation possible and covers a wide range of shapes: from short tail (large p) to long tail (small p) distributions, including the near uniform distribution (very large p), the Gaussian distribution ($p = 2$), and the bilateral exponential distribution $p = 1$. For simplicity, we shall take ψ_i to be a same function ψ.

We first look at the lower bound for the asymptotic variance, which is achieved with the choice $\varphi_p = \text{sign}\,(s)|s|^{p-1}$ for ψ. Note that φ_1 has a jump at the origin, but the result of Section II-B can still be proved by interpreting φ_1' as twice the Dirac function, i.e., $E[\varphi_1'(S)]$ is twice the density of S at the origin. (The proof is much more complex, but basically, one has to show that (2.1') still holds with this interpretation of φ_1'). Fig. 1 plots the above bound times the sample size versus the shape parameter p. It is seen that the bound increases rapidly to infinity when p approaches 2, showing that it is not only impossible to separate Gaussian sources (as is well known) but also very difficult to separate nearly Gaussian ones.

The ratios of the lower bound of the asymptotic variance over the actual asymptotic variance of the estimator is called (asymptotic) efficiency. Fig. 2 plots the efficiencies corresponding to five choices of ψ: φ_4, $\varphi_{1.5}$, $\varphi_4 + c_{PRJ}\varphi_{1.5}$ (with c_{PRJ} chosen by the method of projection of Section II-D), $\varphi_4 + c_{OPT}\varphi_{1.5}$ (with c_{OPT} chosen optimally by trial), and $\varphi_4 + c_1\varphi_2 + c_2\varphi_{1.5}$ (which is obtained by the method of projection, which is optimal). It is seen that the efficiency

$$\mathbf{T} = \begin{bmatrix} |p_1 q_2 - p_2 q_1| + \sqrt{\Delta} & \text{sign}\,(p_1 q_2 - p_2 q_1)\, 2(pq_1 - p_1 q) \\ \text{sign}\,(p_2 q_1 - p_1 q_2)\, 2(pq_2 - p_2 q) & |p_2 q_1 - p_1 q_2| + \sqrt{\Delta} \end{bmatrix} \tag{3.6}$$

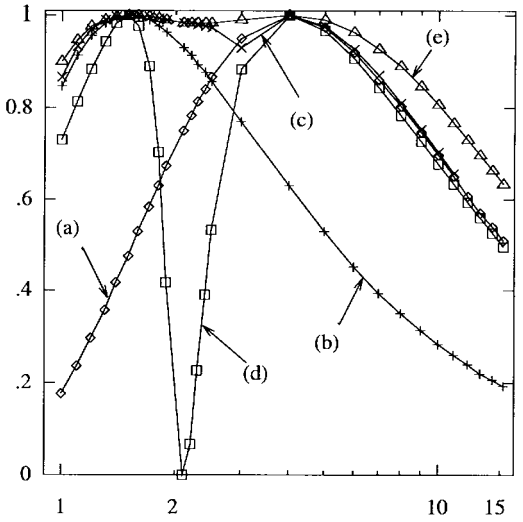

Fig. 2. Asymptotic efficiencies of the estimator for the separating functions. (a) φ_4. (b) $\varphi_{1.5}$. (c) $\varphi_4 + c_{OPT}\varphi_{1.5}$. (d) $\varphi_4 + c_{PRJ}\varphi_{1.5}$. (e) $\varphi_4 + c_1\varphi_2 + c_2\varphi_{1.5}$.

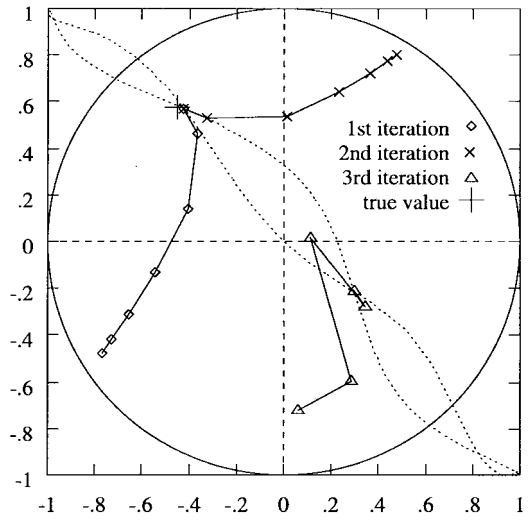

Fig. 3. Examples of iterative computation of the estimator.

for φ_4 is quite good for $p > 2$ (unless p is very large) but poor for $p < 2$. The efficiency for $\varphi_{1.5}$, on the other hand, is good for $p < 2$ but poor for $p > 2$. Thus, one may expect to achieve good efficiency over all by using a linear combination of φ_4 and $\varphi_{1.5}$. This is seen to be the case if the coefficient of the combination is chosen optimally. However, we have to do it by trial and error, as the method of projection of Section II-D is not optimal (the identity function is missing in the projection space). As is seen in Fig. 2, this method can lead to a catastrophic result for p near 2. Since choosing optimally the linear combination of φ_4 and $\varphi_{1.5}$ is impractical and requires the knowledge of the true density of the sources, we are led to consider a linear combination of three functions—φ_4, $\varphi_{1.5}$, and φ_2 (the identity function)—in order to apply the method of projection. The obtained efficiency is even better, approaching 1 for a wide range of value of p.

Note, finally, that the efficiency curves (except for $\varphi_4 + c_{PRJ}\varphi_{1.5}$) seem to be smooth at $p = 2$, even though the asymptotic variance diverges to infinity as $p \to 2$.

B. Off-Line Iterative Algorithm

We have made many simulation experiments to test the iterative (nonadaptive) method of construction of the estimator in Section II-C (for two sources). Fig. 3 illustrates a typical result: Two sources from the same uniform distribution are mixed and then separated using our algorithm with $\phi_1(s) = \phi_1(s) = s^3$ on a block of 128 sample points. We normalize the separating matrix $\mathbf{A}^{-1}$ such that its rows have unit Euclidean norm and represent its off diagonal elements as coordinates of a point in the graph. The two dotted curves represent the sets of solutions of each of the two equations (2.1). Their intersections yield the estimator $\hat{\mathbf{A}}^{-1}$, which, unfortunately, is not unique. Actually, ignoring permutation, there is an $\hat{\mathbf{A}}^{-1}$ lying quite close to the true value $\mathbf{A}^{-1}$ (+ in the graph) and another spurious one. The iteration may converge to the "right" (first and second iterations) or to the "spurious" solution (third iteration), depending on the initial value (it converges well and fast unless one starts at a point on the circle for which the corresponding matrix is singular). To resolve the above ambiguity, one may compute some index of dependence (e.g., the sum of squares of the correlations) between the reconstructed sources and retain the solution for which they are least dependent. In addition, by starting at some good initial estimate, one usually attains the "right" solution.

C. Performance of the Estimator

To study the performance of the estimator, 1000 repeated experiments are performed. In each experiment, two temporally independent signals with identical density are generated, and we apply the method of Section II to separate them. Note that our algorithms are "transformation invariant." If one multiplies the data vectors $\mathbf{X}(t)$ by some matrix $\mathbf{T}$ as well as by the starting point of the algorithm with the same matrix, then the result would also be multiplied by $\mathbf{T}$. Thus, we can (and shall) take as observations the nonmixed signals if we also look at the effect of staring point. Here, we shall assume that we have a good initial estimate of the mixing matrix so that we may disregard the situations where the algorithm converges to a spurious estimate (see Section IV-B). Then, it produces an estimate independent of the starting point and, thus, has a performance independent of the true mixing matrix $\mathbf{A}$. Note that the direct method of Section IV-C3 does not need iteration and, hence, has a performance really independent of $\mathbf{A}$

Two situations are considered in which the sources have the uniform distribution over $[-1, 1]$ and the bilateral exponential distribution, respectively. Fig. 4 plots the square root of the product of the sample size T with the mean squares of δ_{12} and δ_{21} versus T. This quantity should be approximately constant with respect to T, according to the asymptotic analysis. The theoretical asymptotic values are also computed and displayed as dotted horizontal lines. One can see from Fig. 4 the good agreement between the experimental and theoretical results for large T. For the uniform distribution, this is attained fairly quickly, at T of about several hundreds, whereas for the bilateral exponential distribution, this is attained only at about 1000. In both cases, the root mean squares of the

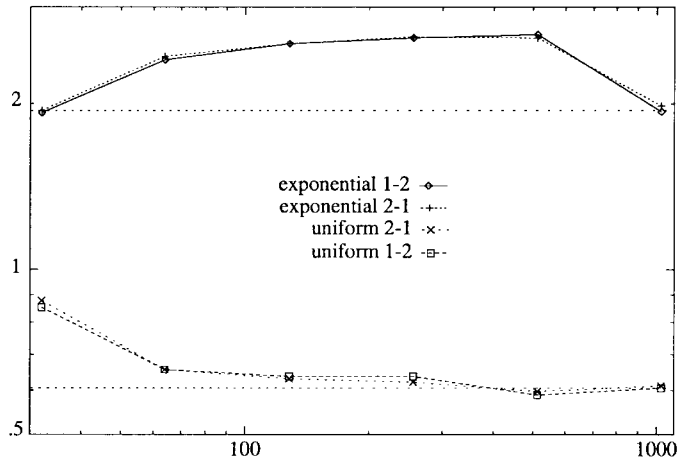

Fig. 4. Simulations results in separating white sources.

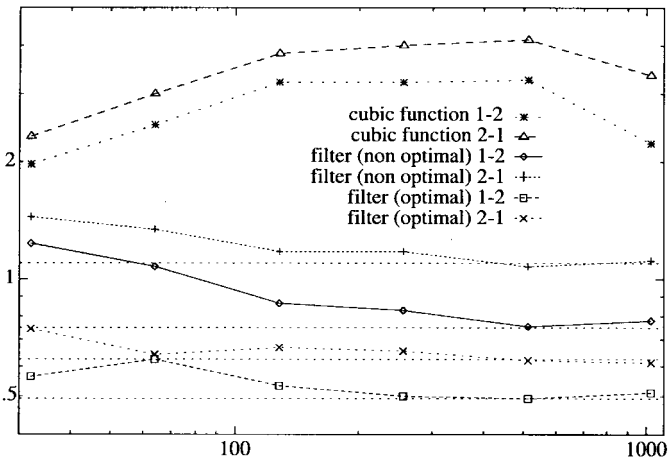

Fig. 5. Simulations results in separating correlated sources.

TABLE I

Methods		Description
1	cubic separating function:	$\psi_1(x) = \psi_2(x) = x^3$
2	linear filter (nonoptimal):	$\phi_{1,k} = 1\ 2\ 1$ $\phi_{2,k} = -1\ 2\ -1$
3	linear filter (optimal):	$\phi_{1,k} = 0.5\ -1.5\ 2.25\ -1.5\ 0.5$ $\phi_{2,k} = -0.6\ 0\ 1.36\ 0\ -0.6$

contamination coefficients also tend to be higher than the asymptotic values for smaller T.

Another simulation experiments concerns the separation of correlated sources. We have generated two signals according to the autoregressive (AR) models:

$$S_1(t) = S_1(t-1) - 0.5 * S_1(t-2) + u_1(t),$$
$$S_2(t) = 0.6 * S_1(t-2) + u_2(t)$$

where u_1 and u_2 are two independent sequences of independent random variables with uniform distribution over $[-1, 1]$. We then applied both methods of Sections II and III to separate them. As before, we shall take the nonmixed signals as observations. Note that the method of Section II is not designed for correlated signals, but we include it to get an idea of how well it can perform. For the method of Section III, we shall use both the optimal filter and nonoptimal linear filter. Table I indicates the methods used and their descriptions:

For speed, the estimators in Methods 2 and 3 are computed by the direct method, based on the simultaneous diagonalization, as described in Section III-C2. Fig. 5 plots the square root of the product of T and the mean squares of δ_{12} and δ_{21} obtained from 1000 repeated experiments versus T. Again, this quantity should be approximately constant with respect to T. The theoretical asymptotic values are also displayed as dotted horizontal lines (there is no simple formula for the asymptotic mean square of the δ_{ij} in the method using separating functions when the sources are correlated). We observe good agreement between the experimental results and the theory attained at fairly moderate T—several hundred. In addition, as in the previous experiment, the mean squares of the contamination coefficients also tend to be higher than the asymptotic values at smaller T.

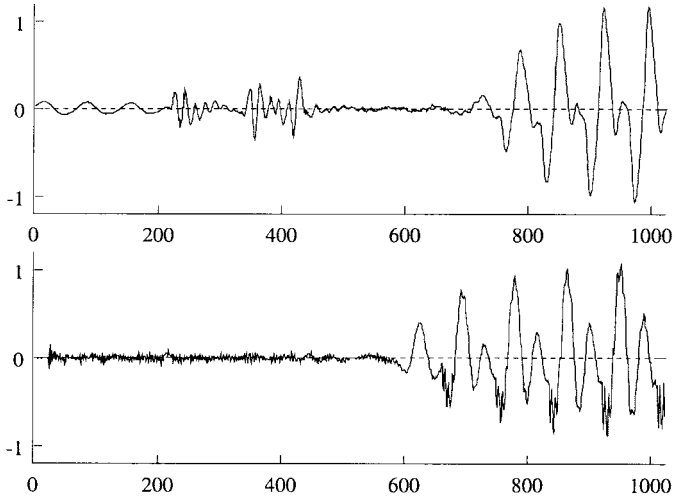

Fig. 6. Original speech signals.

D. A Real Data Example

As a final example, we have performed some experiments with two unrelated segments of real speech signal of length 0.1024 s (1024 samples with sampling frequency 10 kHz). They are artificially mixed, and we apply the method of Section III to separate them. Since the signals are obviously nonstationary, we shall use an adaptive implementation (with direct computation as explained at the end of Section III-C2). The separating filters have impulse responses $1, -2, 3, -2, 1$ and $1, 2, 3, 2, 1$, respectively, and the forgetting factor λ is taken to be 0.999. This yields an effective window size about 1000 points, which is not large since it corresponds only to 0.1 s. Figs. 6 and 7 show the two original speech signals and the reconstructed signals from two mixtures of them, respectively. We do not plot these mixtures since our (direct) method does not depend on them.

It can be seen from Figs. 6 and 7 that after a start-up time about only several dozens of samples, the adaptive algorithm is able to separate the two signals nicely. However, near the end of the record, the two sources become very similar, and then, our algorithm has trouble separating them. This is as expected since the variance of the estimator becomes infinite

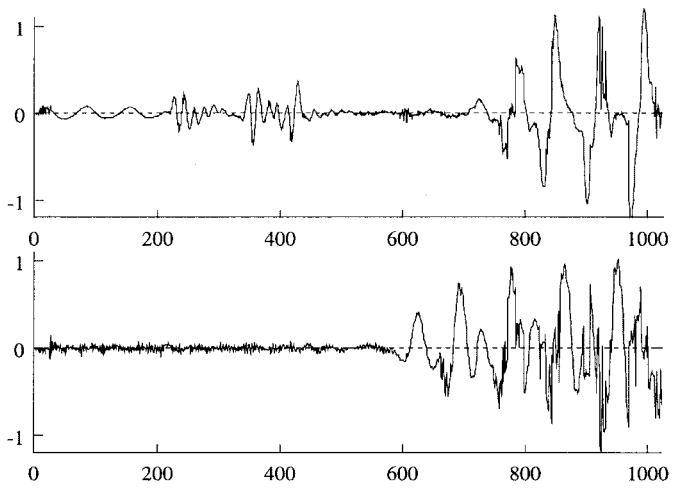

Fig. 7. Reconstructed signals from their mixtures.

when the spectra of the sources are the same (see Section III-B). Nevertheless, in the above example, the separation is still acceptable, and noting that the time span in which the two speech signals have similar spectra is usually quite short, our algorithm performs globally well.

V. Conclusion

We have proposed two methods for separating a mixture of independent sources. They are based on the use of nonlinear separating functions and linear filters. The first method is specially adapted to temporally independent non-Gaussian sources, whereas the second is specially adapted to correlated sources with distinct spectra. A theoretical analysis of the performance of the methods has been made and the choice of the separating function discussed. We have also derived a simple implementation of the second method based on the simultaneous diagonalization of two symmetric matrices. Simulations illustrate the good performance of the methods. Local convergence of the procedure seems quite good; however, the global convergence suffers from the possibility of landing on a "spurious" estimate. This drawback is also present in the similar Jutten–Hérault method. We are currently working toward a way to avoid this.

Acknowledgment

The authors would like to thank C. Jutten and J. Hérault, with whom we have many fruitful discussions, and H. L. N. Thi, who has provided us with the speech data. They would also like to thank several referees who have carefully read our manuscript and suggested many improvements.

References

[1] D. R. Brillinger, *Time Series: Data Analysis and Theory.* New York: Holt, Rinehart, and Winston, 1975.
[2] J.-F. Cardoso, "Source separation using higher order moments," in *Proc. ICASSP 89*, Glasgow, Scotland, May 1989, vol. 4, pp. 2109–2112.
[3] ——, "Iterative technique for blind source separation using only fourth order cumulants," in *Proc. EUSIPCO 92*, Brussels, Belgium, Aug. 1992, vol. 2, pp. 739–742.
[4] J.-F. Cardoso and A. Souloumiac, "An efficient technique for blind separation of complex sources," in *Proc. IEEE SP Workshop Higher-Order Statist.*, Lake Tahoe, CA, 1993, pp. 275–279.
[5] P. Comon, "Independence components analysis, a new concept," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
[6] P. Duvaut, "Principe des méthodes de séparation de sources fondées sur les moments d'ordre supérieur," *Traitement du Signal*, vol. 7, no. 5, (numéro spécial non linéaire non Gaussien), pp. 407–418, 1990.
[7] M. Gaéta and J.-L. Lacoume, "Source separation without *a priori* knowledge: The maximum likelihood solution," in *Proc. EUSIPCO 90*, Barcelona, Spain, 1990, pp. 621–624.
[8] G. H. Golub and F. V. Loan, *Matrix Computations.* Baltimore, MD: Johns Hopkins Univ. Press, 1989.
[9] C. Jutten and J. Hérault, "Blind separation of sources, Part I: An adaptative algorithm based on neuromimetic structure," *Signal Processing*, vol. 24, pp. 1–10, 1991.
[10] J.-L. Lacoume and P. Ruiz, "Source identification: A solution based on the cumulants," in *Proc. 4th ASSP Workshop Spectral Estimation Modeling*, Minneapolis, MN, Aug. 1988, pp. 199–203.
[11] P. Loubaton and N. Delfosse, "Séparation adaptive de sources indépendantes par une approch de déflation," in *Proc. Colloque GRETSI 93*, Juan-les-Pins, France, vol. 1, Sept. 1993, pp. 325–328.
[12] M. Loève, *Probability Theory.* Princeton, NJ: Van Nostrand, 1963.
[13] E. Moreau and O. Macchi, "New self-adaptive algorithm for source separation based on constrast functions," in *Proc. IEEE Signal Processing Workshop Higher-Order Statist.*, Lake Tahoe, CA, 1993, pp. 215–219.
[14] D. T. Pham, P. Garat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *Proc. EUSIPCO 92*, Brussels, Belgium, Aug. 1992, vol. 2, pp. 771–774.
[15] D. T. Pham and P. Garat, "Séparation Aveugle de Sources Temporellement Corrélées," in *Proc. Colloque GRETSI 93*, Juan-les-Pins, France, vol. 1, Sept. 1993, pp. 317–320.
[16] M. Rosenblatt, *Stationary Sequences and Random Fields.* Boston, MA: Birkhauser, 1985.
[17] G. A. F. Seber, *Multivariate Observation.* New York: Wiley, 1984.
[18] R. H. Shumway and Z. A. Der, "Deconvolution of multiple time series," *Technometr.*, vol. 42, pp. 2158–2168, 1995.
[19] L. Tong, V. Soon, Y. F. Huang, and R. Liu, "Indeterminacy and identifiability of blind identification," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 499–509, 1991.
[20] L. Tong, Y. Inouye, and R. Liu, "Waveform preserving blind estimation of multiple independent sources," *IEEE Trans. Signal Processing*, vol. 41, pp. 2461–2470, Oct. 1993.
[21] D. Yellin and E. Weinstein, "Criteria for multichanel signal separation," *IEEE Trans. Signal Processing* vol. 42, pp. 2158–2168, Sept. 1994.

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

He was a Postdoctoral Fellow in the Department of Statistics at the University of California, Berkeley, in 1977 and 1978 and a Visiting Professor in the Department of Mathematics, Indiana University, Bloomington, in 1979 and 1980. He is currently Director of Research at the Centre National de Recherche Scientifique, Joseph Fourier University. His research interests include time series analysis, signal modeling, array processing, biomedical signal processing, and data assimilation in oceanography.

Philippe Garat was born on July 30, 1965 in Talence, France. He received the Engineering Diploma from the Ecole Nationale Supérieure de l'Electronique et de ses Applications, Paris-Cergy, France, in 1988 and the Ph.D. degree in applied mathematics from the Joseph Fourier University, Grenoble, France, in 1994.

He became Assistant Professor at the University of Rennes, Rennes, France, in 1995. His current research interests include computational geometry, statistical modeling, and array signal processing.