OPTIMIZATION OF JADE USING A NOVEL OPTIMALLY WEIGHTED JOINT DIAGONALIZATION APPROACH

Alexander Smekhov and Arie Yeredor

School of Electrical Engineering, Tel-Aviv University
P.O.Box 39040, Tel-Aviv 69978, Israel
Tel: +972 3 6405314; fax: +972 3 6407095
e-mail: {sashas,arie}@eng.tau.ac.il

ABSTRACT

The JADE algorithm (Cardoso and Souloumiac, 1993) is a popular batch-type algorithm for Blind Source Separation (BSS), which employs approximate joint diagonalization (AJD) of fourth-order cumulant matrices, following a whitening stage. In this paper we propose a computationally attractive optimization of JADE for noiseless mixtures, in the form of a post-processing tool. First, we cast the AJD of 4th- and 2nd- order estimated matrices as a weighted least-squares (WLS) problem. We then show (under some commonly met conditions), that in the vicinity of a non-mixing condition (such as at the output of traditional JADE), the asymptotically optimal WLS criterion can be easily formulated and conveniently optimized via a novel algorithm, which uses non-unitary AJD of transformed subsets of the estimated matrices. Optimality with respect to general mixing is maintained, as we show, thanks to the equivariance of the optimal WLS solution. The performance of the new algorithm is analyzed and compared to JADE, identifying the conditions for most pronounced improvement, as demonstrated by simulation.

1. INTRODUCTION

Blind Source Separation (BSS) consists of recovering mixed source signals, possibly via estimation of the mixing matrix $\boldsymbol{A}$ in the (noiseless, invertible) mixture model $\boldsymbol{x}[t] = \boldsymbol{A}\boldsymbol{s}[t], \quad t = 1, 2, \ldots, T$, where $\boldsymbol{s}[t] = \left[s_1[t]\ s_2[t]\ \cdots\ s_N[t]\right]^T$ are N unknown, statistically independent source signals, $\boldsymbol{x}[t] = \left[x_1[t]\ x_2[t]\ \cdots\ x_N[t]\right]^T$ are the observations and $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ is the unknown mixing matrix. The term 'blind' ascribes lack of any additional information regarding the sources or $\boldsymbol{A}$.

One of the most popular batch-type algorithms for BSS is the "Joint Approximate Diagonalization of Eigenmatrices" (JADE, Cardoso and Souloumiac, 1993 [1]). In its pre-processing stage, JADE applies spatial "hard-whitening" to the observations, using the inverse square-root of their empirical correlation matrix, thus causing the transformed observations to be exactly (empirically) spatially white. Following this decorrelation stage, the remaining unitary mixing (rotation) factor is estimated by unitary approximate joint diagonalization (AJD) [2] of a set of fourth-order cumulant matrices, estimated from the whitened observations.

While computationally convenient, the "hard-whitening" stage has been observed ([3, 4, 5]) to limit the attainable performance, as it essentially attributes infinite weight to the fit of the estimated correlation matrix in the AJD process. The exact fit of the diagonalization model to the estimated correlation matrix implies a degraded fit to the estimated cumulants matrices, without proper statistical justification.

In [4] an analysis of the optimal weighting for various combinations of estimated second- and fourth- order cumulants was presented, and the resulting expected performance was outlined. However, the framework of statistic matching in [4] assumes prior knowledge of some of the sources' statistics, on which the optimal weighting depends. Additionally, [4] considers explicitly only subsets of the full set of estimated correlations and cumulants, and no specific algorithm is proposed for the associated nonlinear optimization problem.

In this paper we propose an asymptotically optimal and computationally attractive algorithm for BSS based on the full set of estimated second and fourth order cumulants. The algorithm applies (asymptotically) optimal weighting to the set, and operates by optimally decomposing the full AJD problem into smaller subproblems, conveniently exploiting a recently-proposed tool for non-unitary AJD [5]. Yet, in order to allow tractability of the derivations and of the resulting algorithm, our analysis will be confined to a "non-mixing" ($\boldsymbol{A}$=$\boldsymbol{I}$) condition, showing that optimality is nonetheless maintained with respect to *any* invertible mixing $\boldsymbol{A}$. Consequently, our algorithm would operate as a post-processing tool following any other consistent batch-type BSS algorithm, preferably a 2nd- and 4th- order cumulants based algorithm, such as JADE.

The paper is structured as follows: In the following section we formulate the associated AJD problem and discuss the equivariance of the optimally-weighted AJD solution, thus justifying the selection of a "non-mixing" point for the derivations. In section 3 we consider the asymptotically optimal weighting, leading to construction of the proposed algorithm. Performance analysis and supporting simulation results appear in section 4.

For simplicity, we assume real-valued signals, although the complex-valued case can be similarly considered, but is beyond the scope of this limited-length paper. Additionally, we employ the "working assumption" that all sources have zero-mean and unit variance (see, e.g.,[6] for justification). Two more restrictive, but

nevertheless common assumptions that we make (necessary to establish the asymptotic optimality) are that each source is temporally independent, identically distributed (iid in time) and has zero 3rd-order cumulant.

2. THE APPROXIMATE JOINT DIAGONALIZATION MODEL

We denote by $\boldsymbol{R} = E\left[\boldsymbol{x}[t]\boldsymbol{x}^T[t]\right]$ the observations' correlation matrix, and by $\{\boldsymbol{C}^{kl}\}_{k,l=1,2,...,N}$ their k,l-th cumulant matrix, whose elements are given by

$$C^{kl}_{mn} = cum\{x_k[t], x_l[t], x_m[t], x_n[t]\}_{\ m,n=1,2,...,N}.$$

The respective standard time-average based estimates (see below) will be denoted $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{C}}^{kl}$.

Due to the unit-variance "working assumption" and the multilinearity property of cumulants, we have

$$\boldsymbol{R} = \boldsymbol{A}\boldsymbol{A}^T \quad , \quad \boldsymbol{C}^{kl} = \boldsymbol{A}\boldsymbol{\Lambda}^{kl}\boldsymbol{A}^T \quad _{k,l=1,2,...,N}\ ,$$

where $\{\boldsymbol{\Lambda}^{kl}\}_{k,l=1,2,...,N}$ are, by virtue of the sources' independence, diagonal matrices, with $\Lambda^{kl}_{mm} = A_{km}A_{lm}k_{m\ \ m=1,2,...,N}$, where k_m denotes the fourth-order auto-cumulant (kurtosis) of the m-th source. Thus, a possible estimation approach is to seek the "best fit" of this model to the estimated matrices,

$$\hat{\boldsymbol{R}} \approx \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^T \tag{1a}$$

$$\hat{\boldsymbol{C}}^{kl} \approx \hat{\boldsymbol{A}}\hat{\boldsymbol{\Lambda}}^{kl}\hat{\boldsymbol{A}}^T \quad _{k,l=1,2,...,N}\ . \tag{1b}$$

Two classical approaches to this optimization problem are (i) The pre-whitening based approach, which factors $\hat{\boldsymbol{A}}$ into a non-unitary and a unitary factor, and sets the non-unitary factor so as to attain exact equality in (1a), while adjusting the unitary factor to attain a Least-Squares (LS) fit in (1b); and (ii) The LS approach [5], which attempts to attain an overall LS fit to the entire set. Both of these approaches actually propose a differently-weighted LS (WLS) criterion - however, neither is optimal. It is well-known (e.g., [7]) that under a small-errors assumption, the optimal weighting for WLS (in the sense of minimum mean square estimation error (MMSE)) is the inverse covariance matrix of the "measurements" (left-hand side estimates in (1a,1b)).

It is therefore desirable to use this optimal weighting in fitting (1a,1b). Unfortunately, however, the derivation of optimal weights and solution of the associated WLS problem appear to be quite computationally intractable in the general case. Moreover, since the optimal weights would depend on the unknown sources' statistics (as well as on the unknown mixing), the necessary estimation of these weights would generally introduce further undesirable complexity.

Fortunately, it turns out that the computational burden can be significantly alleviated when assuming a non-mixing condition ($\boldsymbol{A} = \boldsymbol{I}$). Such an assumption is valid (asymptotically) whenever a consistent BSS algorithm is applied to the data prior to applying the optimal WLS approach[1]. However, it is important to show, that the operation of the "pre-processing" BSS algorithm does not undermine the optimality of the WLS solution with respect to the original problem. In other words, while the WLS solution is (asymptotically) optimal for the non-mixing case, we need to show that the combined solution (BSS followed by optimal WLS) maintains the WLS optimality in any (arbitrary) mixing case.

To show that, let us first define the $(N^4+N^2)\times 1$ "measurements vector", comprised of all the estimated cumulants and correlations: $\boldsymbol{y} \triangleq \left[\hat{\boldsymbol{c}}^T \ \ \hat{\boldsymbol{r}}^T\right]^T$ where

$$\hat{\boldsymbol{c}} \triangleq \left[\hat{\boldsymbol{c}}^T_{11}\ \hat{\boldsymbol{c}}^T_{21}\ \cdots\ \hat{\boldsymbol{c}}^T_{12}\ \hat{\boldsymbol{c}}^T_{22}\ \cdots\ \cdots\ \hat{\boldsymbol{c}}^T_{NN}\right]^T$$

with

$$\hat{\boldsymbol{c}}_{kl} \triangleq vec\{\hat{\boldsymbol{C}}^{kl}\}_{\ k,l=1,2,...,N}$$

and $\hat{\boldsymbol{r}} \triangleq vec\{\hat{\boldsymbol{R}}\}$.

Assuming that the correlations and cumulants are estimated using standard time-averaging, namely $\hat{R}_{kl} = \frac{1}{T}\sum_{t=1}^{T} x_k[t]x_l[t]$ and $\hat{C}^{kl}_{mn} = \frac{1}{T}\sum_{t=1}^{T} x_k[t]x_l[t]x_m[t]x_n[t] - \hat{R}_{kl}\hat{R}_{mn} - \hat{R}_{km}\hat{R}_{ln} - \hat{R}_{kn}\hat{R}_{lm}$ (for $k,l,m,n = 1,2,...,N$), it is sufficient to note the following property:

Let $\boldsymbol{v}[t] = \boldsymbol{B}\boldsymbol{x}[t]$ denote some "transformed observations", where $\boldsymbol{B}$ is an arbitrary invertible matrix. Denote $\boldsymbol{y}_{[v]}$ the vector of cumulants and correlations estimated from $\boldsymbol{v}[t]$. Then straightforward multilinear-algebraic manipulations reveal that $\boldsymbol{y}_{[v]}$ and $\boldsymbol{y}$ maintain the linear relation $\boldsymbol{y}_{[v]} = \boldsymbol{W}(\boldsymbol{B})\boldsymbol{y}$, where

$$\boldsymbol{W}(\boldsymbol{B}) \triangleq \begin{bmatrix} \boldsymbol{B}^{\otimes 4} & \boldsymbol{0}_{N^4\times N^2} \\ \boldsymbol{0}_{N^2\times N^4} & \boldsymbol{B}^{\otimes 2} \end{bmatrix}$$

with $\boldsymbol{B}^{\otimes k}$ denoting the k-times Kronecker's product $\boldsymbol{B}^{\otimes k} \triangleq \boldsymbol{B}\otimes\boldsymbol{B}\otimes\cdots\otimes\boldsymbol{B}$ (k terms).

Consequently, mixing (or de-mixing) the observations $\boldsymbol{x}[t]$ with any matrix $\boldsymbol{B}$ merely implies a linear transformation of the "measurements vector" $\boldsymbol{y}$. In the context of a WLS framework, any linear transformation of the measurements is equivalent to a modified weighting, implied by multiplying the original weight matrix on the right and on the left by the transformation matrix and its transpose (respectively). Therefore, for any invertible $\boldsymbol{B}$ (and, subsequently, $\boldsymbol{W}(\boldsymbol{B})$), remixing the observations merely implies an invertible modification of the weight matrix - and therefore applying the *optimal* weights to the remixed observations is equivalent to applying the (generally different) *optimal* weights to the original observations, since, in a sense, the optimal weights can "undo" the remixing effect, if necessary. This property establishes the *equivariance* of the *optimal* WLS solution with respect to the mixing, and enables to apply the optimal WLS criterion at the output of any 'pre-processing" BSS algorithm while maintaining optimality with respect to the original observations.

3. APPLYING ASYMPTOTICALLY OPTIMAL WEIGHTING

The unknown parameters in the WLS model are denoted $\boldsymbol{\theta} \triangleq [\boldsymbol{a}^T \ \ \boldsymbol{\lambda}^T]^T$, where $\boldsymbol{a} \triangleq vec\{\boldsymbol{A}\}$ denotes the

[1] The associated permutation ambiguity is tolerable, since it merely implies different labeling of the sources; Due to the unit-variance "working assumption", we assume no scaling ambiguity.

unknown mixing parameters, and $\boldsymbol{\lambda}$ encompasses all the N^3 unknown diagonal elements in $\{\boldsymbol{\Lambda}^{kl}\}_{k,l=1,2,...,N}$. The desired AJD model can be formulated as $\boldsymbol{y} \approx \boldsymbol{h}(\hat{\boldsymbol{\theta}})$, where $\boldsymbol{h}(\hat{\boldsymbol{\theta}})$ is a nonlinear function of the parameters, prescribed by the applying the $vec(\cdot)$ operator to the right-hand side in (1a,1b).

We now divide $\boldsymbol{\theta}$ into two groups: parameters of interest and nuisance parameters. Note that not all the parameters in $\boldsymbol{a}$ are of interest in our setup: In the vicinity of a non-mixing condition, the resulting Interference to Source Ratio (ISR) depends only on the variance of the estimation error of the *off-diagonal* terms in $\boldsymbol{A}$. More specifically, if we define $\boldsymbol{D} = \hat{\boldsymbol{A}}^{-1}\boldsymbol{A}$ as the resulting overall "contamination matrix", then under the unit-power sources assumption, $\mathrm{ISR}_{ij} \triangleq E[D_{ij}^2]$ is the residual mean contaminating power of source j in the reconstruction of source i. Thus, in the vicinity of a non-mixing condition, it is easy to observe that $\mathrm{ISR}_{ij} \approx E[\hat{A}_{ji}^2]$, so the only parameters of interest (for the overall ISR performance) are the *off-diagonal* elements in $\boldsymbol{A}$. Thus, for attaining optimal performance in terms of ISR, it is sufficient to seek the optimal WLS estimate of these parameters only. We therefore divide $\boldsymbol{\theta}$ into $\boldsymbol{\theta}^{\mathrm{off}}$ and $\boldsymbol{\theta}^{\mathrm{diag}}$, where $\boldsymbol{\theta}^{\mathrm{off}}$ contains onlythe off-diagonal elements of $\boldsymbol{A}$, and $\boldsymbol{\theta}^{\mathrm{diag}}$ contains all nuisance parameters: $\boldsymbol{A}$'s diagonal elements and $\boldsymbol{\lambda}$.

In order to identify which elements of $\boldsymbol{y}$ are relevant for the estimation of $\boldsymbol{\theta}^{\mathrm{off}}$ (under a small-errors assumption), we linearize the model $\boldsymbol{y} \approx \boldsymbol{h}(\boldsymbol{\theta})$ in the vicinity of the true parameters, denoted $\boldsymbol{\theta}_0$. This would also enable a small-errors performance analysis later on. The linearization takes the form

$$\boldsymbol{y} \approx \boldsymbol{h}(\boldsymbol{\theta}) \approx \boldsymbol{h}(\boldsymbol{\theta}_0) + \boldsymbol{H} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

where $\boldsymbol{H}$ denotes the derivative matrix of $\boldsymbol{h}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

We wish to apply model reduction by eliminating all observations (elements of $\boldsymbol{y}$) that are irrelevant to the optimal estimation of $\boldsymbol{\theta}^{\mathrm{off}}$. Note first, that $\boldsymbol{y}$ contains redundancies, due to recurrence of identical values, such as $\hat{\boldsymbol{R}}_{ij} = \hat{\boldsymbol{R}}_{ji}$ and $\hat{\boldsymbol{C}}_{ij}^{ii} = \hat{\boldsymbol{C}}_{ji}^{ii} = \hat{\boldsymbol{C}}_{ii}^{ij} = \hat{\boldsymbol{C}}_{ii}^{ji}$. However, due to the second-order structure of the AJD model, some identical elements are "explained" differently by this model, and therefore not all recurrences can be automatically eliminated. Furthermore, the rules for proper elimination of elements of $\boldsymbol{y}$ must consider not only the model structure, but also the covariance between different elements of $\boldsymbol{y}$. For example, observations that depend only on $\boldsymbol{\theta}^{\mathrm{diag}}$, and are apparently irrelevant to the estimation of $\boldsymbol{\theta}^{\mathrm{off}}$, can only be eliminated if they are uncorrelated with observations that depend on $\boldsymbol{\theta}^{\mathrm{off}}$.

Further details of the elimination process can be found in [8]. It turns out that the reduced linearized model can be broken down into a block-diagonal structure, resulting for all $1 \le i < j \le N$ in:

$$\boldsymbol{y}^{(ij)} \triangleq \begin{bmatrix} \hat{\boldsymbol{C}}_{ij}^{ii} \\ \hat{\boldsymbol{C}}_{ij}^{jj} \\ \hat{\boldsymbol{R}}_{ij} \end{bmatrix} \approx \begin{bmatrix} 0 & k_i \\ k_j & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} A_{ij} \\ A_{ji} \end{bmatrix} \triangleq \boldsymbol{H}^{(ij)}\boldsymbol{\theta}^{(ij)} \tag{2}$$

where $\boldsymbol{y}^{(1,2)}, \boldsymbol{y}^{(1,3)}, ...\boldsymbol{y}^{(N-1,N)}$ are sub-vectors (comprising the reduced observation vector, denoted $\tilde{\boldsymbol{y}}$) which contain the only remaining relevant observations, namely the off-diagonal elements of the estimated cumulant matrices of type $\hat{\boldsymbol{C}}^{ii}$ and of the estimated correlation $\hat{\boldsymbol{R}}$.

Fortunately, the asymptotic covariance analysis (at the non-mixing point) of the reduced observations vector $\tilde{\boldsymbol{y}}$ reveals ([8]) a corresponding block-diagonal structure with 3×3 blocks given by

$$\boldsymbol{C}^{(ij)} \triangleq \mathrm{Cov}\left\{\boldsymbol{y}^{(\mathrm{ij})}\right\} = \frac{1}{\mathrm{T}} \begin{bmatrix} l_i + k_i^2 & k_i k_j & k_i \\ k_j k_i & l_j + k_j^2 & k_j \\ k_i & k_j & 1 \end{bmatrix}$$

where l_i is related to the i-th source's sixth moment via $l_i \triangleq E[s_i^2]E[s_i^6] - E^2[s_i^4] = E[s_i^6] - E^2[s_i^4]$.

An important conclusion is that in the vicinity of $\boldsymbol{A} = \boldsymbol{I}$, optimal weighting can be applied by breaking the WLS problem into the respective set of individual $N(N-1)/2$ WLS sub-problems, applying to each the inverse of the respective covariance matrix (3). Moreover, the unknown $\{k_i\}$ and $\{l_i\}$ can be conveniently estimated from the nearly-separated data using straightforward time-averaging.

One way to apply the required weight matrix to each sub-problem, is to apply a linear transformation $\boldsymbol{Q}^{(ij)}$ to each respective sub-vector $\boldsymbol{y}^{(ij)}$, and then solve an unweighted LS problem. The required transformation is given by the (scaled) inverse square root of $\boldsymbol{C}^{(ij)}$,

$$\boldsymbol{Q}^{(ij)} = \begin{bmatrix} \frac{1}{q_i} & 0 & \frac{k_i}{q_i} \\ 0 & \frac{1}{q_j} & -\frac{k_j}{q_j} \\ 0 & 0 & 1 \end{bmatrix}, \quad {\scriptstyle 1 \le i < j \le N}\ , \tag{3}$$

where $q_i \triangleq \sqrt{l_i}$.

A convenient tool for solving the resulting unweighted LS problem is the AC-DC algorithm [5] for approximate joint diagonalization. Specifically, in view of the required linear transformation (3), we seek to jointly diagonalize the set of three matrices $\hat{\boldsymbol{\mathcal{K}}}^{(ij)}$, $\hat{\boldsymbol{\mathcal{K}}}^{(ji)}$ and $\hat{\boldsymbol{\mathcal{R}}}^{(ij)}$ (for each sub-problem), where

$$\hat{\boldsymbol{\mathcal{K}}}^{(ij)} \triangleq \frac{1}{q_i}\hat{\boldsymbol{\mathcal{C}}}^{(ij)} - \frac{k_i}{q_i}\hat{\boldsymbol{\mathcal{R}}}^{(ij)}\ ,\ \hat{\boldsymbol{\mathcal{K}}}^{(ji)} \triangleq \frac{1}{q_j}\hat{\boldsymbol{\mathcal{C}}}^{(ji)} - \frac{k_j}{q_j}\hat{\boldsymbol{\mathcal{R}}}^{(ij)} \tag{4}$$

$$\hat{\boldsymbol{\mathcal{C}}}^{(ij)} \triangleq \begin{bmatrix} \hat{\boldsymbol{C}}_{ii}^{ii} & \hat{\boldsymbol{C}}_{ij}^{ii} \\ \hat{\boldsymbol{C}}_{ji}^{ii} & \hat{\boldsymbol{C}}_{jj}^{ii} \end{bmatrix} \quad \hat{\boldsymbol{\mathcal{C}}}^{(ji)} \triangleq \begin{bmatrix} \hat{\boldsymbol{C}}_{ii}^{jj} & \hat{\boldsymbol{C}}_{ij}^{jj} \\ \hat{\boldsymbol{C}}_{ji}^{jj} & \hat{\boldsymbol{C}}_{jj}^{jj} \end{bmatrix} \tag{5}$$

$$\hat{\boldsymbol{\mathcal{R}}}^{(ij)} \triangleq \begin{bmatrix} \hat{\boldsymbol{R}}_{ii} & \hat{\boldsymbol{R}}_{ij} \\ \hat{\boldsymbol{R}}_{ji} & \hat{\boldsymbol{R}}_{jj} \end{bmatrix}. \tag{6}$$

The proposed algorithm, given the acronym OFORIA - Optimal Fourth-ORder Identification Algorithm, therefore assumes the following form:

- Apply a consistent BSS algorithm (preferably JADE) to the data $\boldsymbol{x}[t]$. Denote its output $\boldsymbol{v}[t]$.
- Estimate / Calculate (see remark 1 below) $\hat{\boldsymbol{R}}$ and all "central" cumulant matrices $\{\hat{\boldsymbol{C}}^{ii}\}$ of $\boldsymbol{v}[t]$.
- Estimate all $\{\hat{l}_i\}$, extract all $\{\hat{k}_i = \hat{\boldsymbol{C}}^{ii}_{ii}\}$ of $\boldsymbol{v}[t]$.
- For each pair $\{i,j\}_{1\le i<j\le N}$:
  - Construct the transformed matrices triplet $\hat{\boldsymbol{\mathcal{K}}}^{(ij)}, \hat{\boldsymbol{\mathcal{K}}}^{(ji)}, \hat{\boldsymbol{\mathcal{R}}}^{(ij)}$, using $\hat{q}_i = \sqrt{\hat{l}_i}$ and $\hat{k}_i$ for q_i and k_i (resp.) in (4-6);
  - Apply non-unitary, unweighted LS AJD to the triplet, and denote $\hat{\boldsymbol{A}}^{(ij)}$ the 2×2 estimated mixing, resolving scale, permutation and sign ambiguities (see remark 2 below);
  - Place the elements $\hat{A}^{(ij)}_{11}, \hat{A}^{(ij)}_{21}, \hat{A}^{(ij)}_{12}, \hat{A}^{(ij)}_{22}$ in their respective locations $\hat{A}_{ii}, \hat{A}_{ji}, \hat{A}_{ij}, \hat{A}_{jj}$ in the $N\times N$ estimated matrix.
- Denote $\hat{\boldsymbol{A}}_o$ the resulting estimate of the $N\times N$ residual mixing.
- The separated sources are $\hat{\boldsymbol{s}}[t] = (\hat{\boldsymbol{A}}_o)^{-1}\hat{\boldsymbol{v}}[t]$.

Remarks:

1. If JADE is used at the initial stage, and involves estimation of the full set of cumulants, then the cumulants of $\boldsymbol{v}[t]$ can be calculated directly from that set (rather than re-estimated from the data), exploiting the multilinearity of cumulants (and their estimates) as discussed in section 1. Additionally, due to the hard-whitening applied by JADE, the empirical correlation matrix of $\boldsymbol{v}[t]$ is known to be $\boldsymbol{I}$, and needs not be re-estimated either.
2. The scale, permutation and sign ambiguities in $\hat{\boldsymbol{A}}_{ij}$ are resolved by ordering and rescaling its columns such that its diagonal is all-ones and contains the leading elements.

4. PERFORMANCE

Under the small-errors assumption, the covariance in the estimation of the elements of $\boldsymbol{A}$, and hence the ISR, can be predicted analytically from the linearized model and the known covariance matrix of $\tilde{\boldsymbol{y}}$. Exploiting the block-diagonal structure (at $\boldsymbol{A}=\boldsymbol{I}$), we obtain

$$Cov\{\hat{\boldsymbol{\theta}}^{(ij)}\} = \left[\boldsymbol{H}^{(ij)T}\boldsymbol{C}^{(ij)-1}\boldsymbol{H}^{(ij)}\right]^{-1} = \frac{1}{T}\begin{bmatrix} 1+k_i^2/l_i^2 & 1 \\ 1 & 1+k_j^2/l_j^2 \end{bmatrix}^{-1}$$

$$\mathrm{ISR}_{ij}^{\mathrm{opt}} = \mathrm{Var}\{\hat{A}_{ij}\} = \frac{1}{T}\cdot\frac{l_il_j + l_ik_j^2}{l_ik_j^2 + l_jk_i^2 + k_i^2k_j^2} \qquad (7)$$

as the optimal ISR attained asymptotically by the optimal WLS solution at $\boldsymbol{A}=\boldsymbol{I}$, as well as at any other $\boldsymbol{A}$, due to the equivariance of the optimal WLS solution as discussed in Section 1. Likewise, the performance of JADE can be shown ([3],[8]) to be given by

$$\mathrm{ISR}_{ij}^{\mathrm{JADE}} = \frac{1}{T}\cdot\frac{k_j^4 + l_ik_i^2 + l_jk_j^2}{(k_i^2+k_j^2)^2}. \qquad (8)$$

The most pronounced improvement in ISR_{ij} is attained by OFORIA when sources i,j are both sub-Gaussian (with small l_i, l_j), or when source i is sub-Gaussian and source j is super-Gaussian. The advantage of OFORIA is less significant when both sources are strongly super-Gaussian, and is negligible when either of them tends to Gaussianity (or to any other null-kurtic distribution). Obviously, however, OFORIA is always (asymptotically) at least as good as JADE, due to its optimality.

To demonstrate the performance improvement (and support the analysis) we present simulation results for the following setup: $N=4$ sources were mixed using randomly generated[2] mixing matrices $\boldsymbol{A}$. The sources' distributions (all zero-mean, unit-variance) and their respective values of k_i and l_i were:

1. Uniform: between $-\sqrt{3}$ and $\sqrt{3}$; $k_1=-1.2$, $l_1\approx 0.6171$ (sub-Gaussian);
2. Laplace: double-sided exponential with parameter $\sqrt{2}$; $k_2=3$, $l_2=6$ (super-Gaussian);
3. Standard Gaussian: $k_3=0$, $l_3=6$ (Gaussian);
4. Gaussian mixture (GM): two equi-probable Gaussians with means ± 1 and variance 0.2 (down-scaled by $\sqrt{1.04}$ to maintain unit variance); $k_4\approx -1.8491$, $l_4\approx 0.1644$ (sub-Gaussian).

JADE was applied to the mixed data of length $T=1000$, followed by application of OFORIA. Empirical ISR results are summarized in Table 1, along with the analytically predicted values.

		Uniform	Laplace	Gaussian	GM
Uniform	J	-	-23 (-23)	-32 (-34)	-32 (-33)
	O	-	-33 (-34)	-33 (-34)	-35 (-35)
Laplace	J	-25 (-23)	-	-23 (-22)	-26 (-25)
	O	-30 (-29)	-	-23 (-22)	-31 (-30)
Gauss.	J	-28 (-28)	-22 (-22)	-	-30 (-30)
	O	-28 (-28)	-22 (-22)	-	-30 (-30)
GM	J	-38 (-38)	-25 (-24)	-40 (-43)	-
	O	-43 (-43)	-43 (-43)	-43 (-43)	-

Table 1: ISR results [dB] for JADE (J) and OFORIA (O), averaged over 1000 trials. Theoretically predicted values from (7), (8) appear in parentheses.

REFERENCES

[1] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE - Proceedings -F*, vol. 140, no. 6, pp. 362–370, 1993.

[2] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 1, pp. 161–164, 1996.

[3] J.-F. Cardoso, "On the performance of orthogonal source separation algorithms," *Proc. EUSIPCO'94*, pp. 776–779, 1994.

[4] J.-F. Cardoso, S. Bose, and B. Friedlander, "On optimal source separation based on second and fourth order cumulants," *Proc. IEEE Workshop on SSAP*, pp. 198–201, 1996.

[5] A. Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1545–1553, 2002.

[6] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[7] S. M. Kay, *Fundamentals of Statistical Signal Processing*, Prentice-Hall, 1993.

[8] A. Smekhov, "Asymptotically optimized blind source separation based on fourth- and second-order statistics," M.S. thesis, Tel-Aviv University, 2004.

[2] Elements independently redrawn (per each trial) from a standard Normal distribution.