Multiple-Snapshots BSS with General Covariance Structures: a Partial Maximum Likelihood Approach Involving Weighted Joint Diagonalization

Arie Yeredor

School of Electrical Engineering, Tel-Aviv University

Abstract

Maximum Likelihood (ML) blind separation of Gaussian sources with different temporal covariance structures generally requires the estimation of the underlying temporal covariance matrices. The possible availability of multiple realizations ("snapshots") of the mixtures (all synchronized to some external stimulus) may enable such estimation. In general, however, since these temporal covariance matrices are high-dimensional, reliable estimation thereof might require a prohibitively large number of snapshots. In this work, we propose to take an alternative, partial and approximate ML approach, which regards a selected set of spatial sample-generalized-correlations of the observations (rather than the observations themselves) as the “front-end” data for the ML estimate. As we show, the implied Correlations-Based approximate ML (CBML) estimate, which can also be regarded as a weighted joint diagonalization approach, requires the estimation of considerably smaller covariance matrices, and can therefore be preferable to the “full” Data-Based ML (DBML) estimate. Therefore, although asymptotically sub-optimal, under sub-asymptotic conditions CBML can outperform the asymptotically-optimal DBML, as we demonstrate in simulation.

Keywords: ICA, BSS, Maximum Likelihood separation, weighting.

Email: arie@eng.tau.ac.il, Tel.:(+972)-3-6405314, Fax:(+972)-3-6407095.

A preliminary, partial version of this work will be presented at ICASSP’2011 [16].

Preprint submitted to Signal Processing July 13, 2011
1. Introduction

Consider the classical real-valued, linear, invertible, static, and noiseless mixture-model $X = AS$, where $S \triangleq [s_1, s_2, \ldots, s_K]^T \in \mathbb{R}^{K \times N}$ is a matrix containing the $K$ unobserved, zero mean, mutually statistically independent source signals (each of length $N$) as its rows; $A \in \mathbb{R}^{K \times K}$ is the unknown mixing matrix (assumed to be nonsingular); and $X \triangleq [x_1, x_2, \ldots, x_K]^T \in \mathbb{R}^{K \times N}$ is the matrix of $K$ observed mixture signals. Assume further, that each source has its own (generally unknown) temporal covariance matrix

$$C_k \triangleq E[s_k s_k^T] \in \mathbb{R}^{N \times N}. \quad (1)$$

While these covariance matrices may take some particular structures (e.g., a Toeplitz structure for stationary sources), we do not assume any particular structure in this work, and therefore general, not necessarily stationary sources are assumed. We do assume, however, that all $C_k$ are positive-definite (namely, nonsingular).

Although second-order statistics (SOS) are generally insufficient for consistent separation, if the sources exhibit diverse temporal covariance structures, consistent separation can indeed rely exclusively on SOS. Moreover, if the sources are Gaussian, derivation of the Maximum Likelihood (ML) separation with respect to the observed data becomes tractable.

Prior work on ML separation was mainly limited to a few important, yet merely particular cases in terms of the sources’ temporal covariance structures. The most prominent is the case of stationary sources, addressed, e.g., by Pham and Garat [10] for general stationary sources and by Dégerine and Zaïdi [4] for Auto-Regressive (AR) sources. Another particular case, involving nonstationary structures with varying variance-profiles, was considered by Pham and Cardoso [9]. ML separation of cyclostationary sources was also considered, e.g., by Pham in [7].

Recently, ML separation and bounds on the attainable Interference to Source Ratio (ISR) were derived in [15] for the more general real-valued case sources.
with arbitrary temporal covariance matrices $C_k$, not structured in any particular way. In a “semi-blind” scenario, when these covariance matrices are known, they are used directly for the ML estimation (see more details in Section 2). However, in the more realistic “fully-blind” scenario, when the $C_k$ are unknown, they need to be estimated (together with the mixing or demixing matrix) from the observed data. To this end, we assume that multiple realizations (“snapshots”) of the mixture are available.

In a “multiple-snapshots” scenario several (say $M$) snapshots of statistically-independent realizations of the same mixture are given,

$$X^{(m)} = AS^{(m)}, \ m = 1, ..., M. \ (2)$$

It is further assumed that each of these independent snapshots is synchronized to some external “trigger” or stimulus, which governs the temporal structures of the sources, such that the $k$-th source in each snapshot can be assumed to be independently drawn from the same zero-mean distribution with covariance $C_k$. Note that the mixing-matrix is assumed to remain constant along the different snapshots.

Generally, the estimation of a covariance matrix from a single, or even from several snapshots results in poor accuracy, unless the matrix can be succinctly parameterized (e.g., in the case of parametric stationary sources). In the most general case, the number of free parameters describing an unstructured $N \times N$ covariance matrix is $\frac{1}{2}N(N+1)$ (in our case we can drop one parameter from this total count, due to the tolerable scaling ambiguity), and therefore the number of independent realizations required for reliable estimation thereof might be too huge for practical consideration.

An alternative approach to using ML estimation with respect to (w.r.t.) the observed mixtures can be the use of ML estimation w.r.t. a selected set of $P$ spatial $K \times K$ sample-generalized-correlation (SGC) matrices (possibly, but not necessarily, lagged correlation matrices at different lags). A general expression for such an SGC matrix takes the form

$$\hat{R} = XPX^T \in \mathbb{R}^{K \times K}, \ (3)$$
where $P \in \mathbb{R}^{N \times N}$ is some selected “association matrix” (see also [14]). Different association matrices give rise to different types of spatial sample-correlation matrices. For example:

- If $P = \frac{1}{N} I$ (where $I$ denotes the Identity matrix), then $\hat{R}$ is the spatial zero-lag sample-correlation of the observations;

- If $P$ is taken as an all-zeros matrix with $\frac{1}{2(N-|\ell|)}$ along its $\pm \ell$-th diagonals, $\hat{R}$ becomes the unbiased, symmetrized estimate of the observations’ lagged correlation matrix at lag $\ell$;

- Likewise, if $P$ is a general Toeplitz matrix, $\hat{R}$ can be regarded as some linear combination of estimated correlation matrices at different lags;

- If $P$ is taken as an all-zeros matrix with a sequence of $Q$ non-zero values of $\frac{1}{Q}$ somewhere along its main diagonal, $\hat{R}$ becomes the spatial zero-lag sample-correlation taken over the respective time-segment of length $Q$;

- Spectral matrices at certain frequencies (e.g., as used in [10]), time-frequency matrices at certain time-frequency points (e.g., as used in [8]), or cyclic correlation matrices (e.g., as used in [7]) can also be obtained by setting $P$ to the appropriate (possibly complex-valued) transformation matrices.

No matter how $P$ is chosen, such a matrix $\hat{R}$ always satisfies

$$\hat{R} = XPX^T = A \left( SPS^T \right) A^T, \quad (4)$$

and therefore

$$E \left[ \hat{R} \right] = AE \left[ SPS^T \right] A^T. \quad (5)$$

Fortunately, due to the mutual statistical independence of the sources, the $(k, \ell)$-th element of $E \left[ SPS^T \right]$, which equals $E[s_k^T P s_{\ell}]$ is zero for all $k \neq \ell$, so $E \left[ SPS^T \right]$ is always diagonal, regardless of the particular choice of $P$.

Thus, given a selected set of some $P$ symmetric association matrices $P_1, \ldots, P_P$, it is possible to construct a respective set of $P$ spatial SGC matrices, $\hat{R}_p =$
\( \mathbf{X} \mathbf{P}_p \mathbf{X}^T \) (for \( p = 1, \ldots, P \)), and the mixing matrix \( \mathbf{A} \) can then be estimated from the Approximate Joint Diagonalization (AJD) of these matrices.

In fact, AJD of such matrices (taken from a single snapshot) has been used in classical BSS methods such as SOBI [1], but not in a ML framework. Recognizing that (due to the central limit theorem) these sample-correlations become asymptotically jointly Gaussian (as long as the association matrices are sufficiently “rich” and all sources have finite fourth-order moments and satisfy the so-called “mixing condition” - e.g., [3]), it can be shown that (asymptotically) ML estimation w.r.t. these sample-correlation matrices takes the form of a weighted AJD.

Moreover, when the sources are nearly separated (e.g., following some initial sub-optimal separation stage), the optimal weight matrix becomes (nearly) block-diagonal (with \( \frac{1}{2} K (K-1) \) blocks, each \( P \times P \), see further details in Section 3). In order to estimate the optimal weights for such weighting, it is sufficient to estimate the respective \( P \times P \) covariance matrices, for which a modest number of snapshots can be sufficient.

We therefore propose in here to substitute the direct Data-Based ML (DBML) approach (w.r.t. to the observed mixtures), which requires estimation of \( K \) full \( N \times N \) covariance matrices, with a Correlations-Based approximate ML (CBML) approach (w.r.t. a selected set of \( P \) spatial SGC matrices), which merely requires the estimation of \( \frac{1}{2} K (K-1) \) much smaller \( P \times P \) covariance matrices.

In the special case where the sources are all stationary Gaussian AR sources of maximal order \( P - 1 \), the \( P \) lagged sample correlation matrices (at lags 0 to \( P - 1 \)) form (asymptotically) sufficient statistics w.r.t. the mixture, and therefore in this particular case, both methods are asymptotically equivalent (and optimal), assuming that the known AR structure is exploited both for the estimation of the full (Toeplitz) covariance matrices in the first approach and for the smaller covariance matrices in the second (see [11]). However, in other, more general cases, the CBML approach is obviously sub-optimal, and yet with a small number of available snapshots, this sub-optimal approach can significantly outperform the DBML approach, which is only asymptotically optimal.
The paper is structured as follows. In the next section we briefly review the “full” DBML approach. In Section 3, we present the proposed CBML approach, together with asymptotic performance analysis. Simulation results are presented in Section 4, demonstrating the advantages of CBML under sub-asymptotic conditions. Our concluding remarks are summarized in Section 5.

2. Data-Based ML

In this section we provide a brief overview of the DBML approach (already considered in [15]), assuming that all the source signals are Gaussian. For simplicity, we begin with the “semi-blind” scenario, in which all $C_k$ are known, and then consider the more realistic, “fully-blind” case, in which they are unknown.

As shown in [15], in the semi-blind case the ML estimate of the demixing matrix $B \triangleq A^{-1}$ can be obtained (from a single snapshot) as follows:

1. Compute the following “target-matrices” (for $k = 1, ..., K$)

$$\hat{R}_k = \frac{1}{N} X C_k^{-1} X^T \in \mathbb{R}^{K \times K}; \quad (6)$$

2. Find the matrix $\hat{B}$, such that the $K$ transformed matrices $\hat{R}_k = \hat{B} \hat{R}_k \hat{B}^T$ each satisfies

$$\hat{R}_k e_k = e_k, \quad k = 1, ..., K,$$

$e_k$ being the $k$-th column of the $K \times K$ Identity matrix.

Since the sources’ covariance matrices, $C_k$, were assumed known, they do not have to be estimated from the data, and therefore a single snapshot is sufficient in this case. Still, if several (independent) snapshots are available, the DBML estimate takes a very similar form, the only difference being that the target matrices $\hat{R}_k$ are replaced with their empirical average over snapshots, namely

$$\hat{R}_k = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{N} X^{(m)} C_k^{-1} (X^{(m)})^T \quad k = 1, ..., K. \quad (8)$$

Note that the set of equations (7) on the second step above amounts to a specially-structured form of joint diagonalization, which we term ‘Hybrid Exact-Approximate joint Diagonalization” (HEAD), since this set essentially requires
that each \((k\text{-th})\) matrix of the \(K\) transformed matrices \(\hat{R}_k\) be “exactly diagonal” in its \(k\)-th column (and row), whereas all its other elements are irrelevant. The HEAD problem is discussed in detail in [13], where several iterative solution algorithms are outlined. For example, one of the most simple solutions (which we briefly outline in here for completeness), proposed by Dégerine and Zaïdi [4], consists of alternating updates of the rows \(b_k^T\) of \(\hat{B}\) as follows. Given some initial guess of \(\hat{B}\) (possibly the Identity matrix), repeat the following “sweep” for \(k = 1, ..., K\):

\[
\hat{b}_k' \leftarrow \hat{b}_k - \hat{B}^T_{(k)} \left( \hat{B}_{(k)} \hat{R}_k \hat{B}^T_{(k)} \right)^{-1} \hat{B}_{(k)} \hat{R}^{(k)} \hat{b}_k
\]

\[
\hat{b}_k \leftarrow \frac{\hat{b}_k'}{\sqrt{\hat{b}_k'^T \hat{R}_k \hat{b}_k'}},
\]

(9)

where \(\hat{B}_{(k)} \in \mathbb{R}^{(K-1) \times K}\) denotes the matrix \(\hat{B}\) without its \(k\)-th row. When such “sweeps” are repeated iteratively, convergence to a solution of (7) is usually attained within several iterations (see [4] or [13] for more details).

In the fully-blind scenario the sources’ covariance matrices \(C_k\) are unknown, and therefore have to be estimated from the data together with \(B\). In this case the exact ML estimate calls for joint maximization of the likelihood w.r.t. both \(B\) and all \(C_k\). Nevertheless, a good approximation of the ML covariance matrices’ estimates can be obtained by first applying any consistent\footnote{By “consistent” we generally mean “\(\sqrt{N}\)-consistent”, namely an algorithm for which the root-mean-square estimation error converges to zero as \(1/\sqrt{N}\).} BSS algorithm to the mixtures and then estimating each covariance matrix \(C_k\) from the respective estimated source \(\hat{s}_k\).

Although in several particular cases (e.g., parametric stationary or cyclo-stationary ergodic signals) the covariance can be consistently estimated from a single (sufficiently long) realization of the mixture signals, the general case calls for the use of several snapshots.

The DBML estimation process for the “fully-blind” case therefore proceeds as follows.

1. Apply some initial separation, using any consistent (albeit sub-optimal)
BSS algorithm, e.g., SOBI\(^2\) [1] (preferably, use correlation estimates averaged over all available snapshots). Denote the resulting demixing matrix \(\hat{B}_0\) and obtain the estimates of the sources from all snapshots \(m = 1, \ldots, M\),

\[
\hat{S}^{(m)} \triangleq [\hat{s}_1^{(m)} \ldots \hat{s}_K^{(m)}]^T = \hat{B}_0 \cdot X^{(m)}.
\]

(10)

2. Use the estimated sources to obtain the following sample-covariance estimates of all \(C_k\) (which would be consistent in \(M \to \infty\), but singular for \(M < N\); alternative (non-ML) estimates, e.g., [6, 2] may be considered as well):

\[
\hat{C}_k = \frac{1}{M} \sum_{m=1}^{M} \hat{s}_k^{(m)} \hat{s}_k^{(m)\, T} \quad k = 1, \ldots, K.
\]

(11)

3. Compute the “target-matrices” (for \(k = 1, \ldots, K\)), substituting (in (8)) the original observations with the nearly-separated sources and the true (unknown) covariance matrices with their (approximate) ML estimates:

\[
\hat{R}_k = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{N} \hat{S}^{(m)} \hat{C}_k^{-1} \hat{S}^{(m)\, T} \quad k = 1, \ldots, K.
\]

(12)

4. Find the matrix \(\hat{B}_1\), which solves the HEAD problem (7) for \(\hat{R}_1, \ldots, \hat{R}_K\).

The overall (approximate) ML estimate of the demixing matrix is \(\hat{B} = \hat{B}_1 \cdot \hat{B}_0\).

If desired, several additional iterations can be applied, assuming that the better-separated sources can offer a better estimate of their respective covariance matrices. In fact, such an iterative scheme is equivalent to alternating-coordinates maximization of the likelihood, alternating between maximization w.r.t. \(B\) and w.r.t. all the covariance matrices \(C_k\). Although such a scheme is generally not guaranteed to converge to the true global maximum of the likelihood function, it usually does.

Note that the use of the ML sample-covariance estimates (11) in (12) is impossible for \(M < N\), since in that case the resulting \(\hat{C}_k\) are singular (and

\(^2\)For stationary, or nearly-stationary sources, the consistency of SOBI is guaranteed; In general, however, SOBI may not be consistent, and some alternative initial separation method should be used.
are no longer the ML estimates), and cannot be inverted; however, other covariance estimates, e.g., [6, 2], may be used - at the expense of deviation from the true ML estimate. Nevertheless, as long as such approximate ML estimates of the covariance matrices are consistent (converge to the true matrices as $M \to \infty$), the resulting (approximate) ML estimate of the demixing matrix $B$ is asymptotically efficient (attains the Cramér-Rao bound as $M \to \infty$) - even if the estimates of the $C_k$ are not exactly their ML estimates. This property is associated with the observation [15] that the Fisher Information Matrix with respect to the demixing matrix’ elements and the covariance matrices’ elements is block-diagonal, such that the two groups are decoupled.

3. Correlations-Based ML

While the full ML approach relies on the specially-structured joint diagonalization ("HEAD") of the particular “target-matrices” $\hat{R}_1, \ldots, \hat{R}_k$ in (6) or (8), some well-known classical (non-ML) methods rely on “conventional” AJD of sample-correlation matrices at different lags (e.g., SOBI [1]), or of zero-lag sample-correlation matrices taken over different sub-intervals of the observation interval (e.g., BGL [9]).

As already mentioned above, any such spatial SGC matrix can be expressed in the general form $\hat{R} = XPX^T$, where $P \in \mathbb{R}^{N \times N}$ is a symmetric association matrix.

Therefore, the ML “target matrices” $\hat{R}_1, \ldots, \hat{R}_k$ are merely particular cases of such SGC matrices, obtained with the specific association matrices $P_k = \frac{1}{N}C_k^{-1}$ - which require knowledge (or a good estimate) of all $C_k$. Suppose now, that the sources’ temporal covariance matrices are unknown and cannot be estimated to within acceptable accuracy from the available snapshots. In other words, due to the large total number of $\frac{1}{2}KN(N+1) + K^2$ unknown parameters (or $\frac{1}{2}KN(N+1) + K(K-1)$, considering the tolerable scale ambiguity), the number $MKN$ of measurements is too small for the ML estimate to assume its asymptotic efficiency properties. It is then conceivable to consider some
alternative, heuristically-chosen set of $P$ association matrices $P_1, \ldots, P_P$ (where $P$ does not even have to equal $K$), which do not require knowledge of the sources’ covariance matrices $C_k$. For example, to obtain the set of symmetric lagged correlation matrices at lags $0, 1, \ldots, P-1$ (for some selected number of lags $P$), the respective set of association matrices $P_1, \ldots, P_P$ would be chosen such that $P_1 = \frac{1}{P} I$, and each of the other $P_p$ is a symmetric Toeplitz matrix, all zeros except for the value $\frac{1}{2} \left( \frac{1}{P} \right)$ along both its $(p-1)$-th sub-diagonal and super-diagonal.

We can now regard this (or any other) set of $P$ spatial SGC matrices as the new “front-end” data, and seek a “Correlations-Based” approximate ML estimate w.r.t. these matrices (rather than seek the “full” Data-Based ML estimate, w.r.t. the original data, which requires the estimation of the large covariance matrices). Of course, unless these SGC matrices form a sufficient statistic w.r.t. the data, the resulting CBML estimate would be asymptotically inferior to the full DBML estimate. Nevertheless, under more realistic, sub-asymptotic conditions (and in a fully-blind scenario) CBML can generally outperform DBML, since it essentially requires the estimation of considerably fewer parameters: as we shall see, instead of requiring the estimation of $K$ covariance matrices of dimension $N \times N$ (in addition to the $K^2$ elements of the demixing matrix), it merely requires the estimation of $\frac{1}{2} K(K-1)$ covariance matrices of much smaller dimensions $P \times P$. With $P, K << N$, the difference in the resulting accuracy of the demixing matrix’ estimate becomes rather significant, as we shall show in the sequel.

As already proposed for the fully-blind scenario, we assume that some initial separation algorithm has been applied to the original observations, so that the observed mixture now consists of the “nearly separated” sources $\hat{S}^{(m)}$ in each snapshot. Note that this assumption merely requires a ”sufficiently good” separation algorithm for the signals at hand. For example if the signals are known to be nearly stationary, one may use the SOBI [1] or WASOBI [12, 11] methods, which would provide reasonable initial separation, to be later refined by using our proposed CBML method. Also, if one has a particular set of SGC matrices
suited for the signals, then initial separation using ordinary joint diagonalization of this set may also be sufficient for subsequent application of CBML to the same set.

Thus, assume some selected set $P_1, \ldots, P_P$ of $P$ association matrices, and let the implied SGC matrices in each snapshot ($m = 1, \ldots, M$) be given by

$$
\hat{R}_p^{(m)} = S^{(m)} P_p (S^{(m)})^T \in \mathbb{R}^{K \times K}, \quad p = 1, \ldots, P,
$$

(13)

An appealing property of these $\hat{R}_p^{(m)}$ is that, as an extension of the Central Limit Theorem, under some mild conditions on the sources’ distributions, as well as on the “richness” and on the condition number of the association matrices $P_p$, they would usually all be nearly jointly-Gaussian (in each snapshot) for sufficiently large $N$, and statistically independent between snapshots:

1. Within each snapshot $m$, each element of $\hat{R}_p^{(m)}$ is given by a linear combination of the form:

$$
\hat{R}_p^{(m)}[k, \ell] = \sum_{n_1, n_2=1}^{N} P_p[n_1, n_2] \cdot \hat{s}_k^{(m)}[n_1] \cdot \hat{s}_\ell^{(m)}[n_2],
$$

(14)

where $\hat{R}_p^{(m)}[k, \ell]$ denotes the $(k, \ell)$-th element of the matrix $\hat{R}_p^{(m)}$, and where $\hat{s}_k^{(m)}[n_1]$ denotes the $n_1$-th element (sample) of the $k$-th estimated source signal $\hat{s}_k^{(m)}$ (with similar interpretations of $P_p[n_1, n_2]$ and $\hat{s}_\ell^{(m)}[n_2]$).

When $N$ is large and $P_p$ is sufficiently rich, the number of elements in this sum is very large, and therefore mild conditions (sometimes termed “mixing conditions”, see, e.g., [3]) on the distributions of the elements $\hat{s}_k^{(m)}[n_1]$, $\hat{s}_\ell^{(m)}[n_2]$ are sufficient for this sum to have a nearly-Gaussian distribution (regardless of the specific underlying distributions of the sources). Moreover, since the total number of elements in each sample-correlation matrix is much smaller than $N$ (and fixed, namely does not grow with $N$), all of these elements are also (nearly) jointly Gaussian.

2. Since the snapshots are assumed to be statistically independent, sample-correlation matrices taken from different snapshots are statistically independent as well.
Thus, since all the elements of these matrices are (approximately) jointly Gaussian, we only need to find their mean and covariance in order to be able to specify their (approximate) joint distribution and the ensuing CBML estimate - the (approximate) ML estimate of $B$ based on these matrices alone. For convenience of the exposition, let us first consider a single-snapshot (and therefore temporarily drop the snapshot-index $m$). In addition, let us rearrange the elements of these matrices (taken in a single snapshot) in a particular form of new vectors as follows. First, we define (for each $(k, \ell) \in [1, K]$) the vectors

\[
\hat{r}_{k,\ell} \triangleq \begin{bmatrix} \hat{R}_1[k,\ell] & \cdots & \hat{R}_P[k,\ell] \end{bmatrix}^T \in \mathbb{R}^P,
\]

where $\hat{R}_p[k,\ell]$ denotes the $(k, \ell)$-th element of $\hat{R}_p$. Then, these vectors can be concatenated into two groups. One group consists of all the vectors which contain diagonal terms of the matrices, and would therefore be called the Diagonal-Terms (DIT) vector,

\[
\hat{q} \triangleq \begin{bmatrix} \hat{r}_1^T & \hat{r}_2^T & \cdots & \hat{r}_K^T \end{bmatrix}^T \in \mathbb{R}^{PK}.
\]

The second group consists of all the other vectors, but since all the target-matrices are symmetric, it is sufficient to only account for $\hat{r}_{k,\ell}$ with $k < \ell$,

\[
\hat{r} \triangleq \begin{bmatrix} \hat{r}_{1,2}^T & \hat{r}_{1,3}^T & \cdots & \hat{r}_{K-1,K}^T \end{bmatrix}^T \in \mathbb{R}^{\frac{1}{2}PK(K-1)}.
\]

We call this vector the “Off-Diagonal Terms” (ODIT) vector (see Figure 1 for an illustration of this vector for the case $K = 3, P = 4$). The “front-end” data for the CBML estimate consists of the DIT and ODIT vectors taken in each snapshot. We now need expressions for the means and for the joint covariance of these vectors.

Figure 1: (appended at the end) Construction of the ODIT vector, $K = 3, P = 4$.

Since we assume that the initial separation stage brought the observations to a nearly-separated state, we may represent these new observations (omitting
the snapshot-index) as a new mixture

$$\hat{S} = (I + \mathcal{E})S$$  \hspace{1cm} (18)

of the original sources, where $(I + \mathcal{E})$ is the “residual mixing” (nearly non-mixing) matrix, such that $\mathcal{E} \in \mathbb{R}^{K \times K}$ is a matrix with very small off-diagonal elements and an all-zeros diagonal. Thus, for each generalized sample-correlation matrix $\hat{R}_p$ we have (for $p = 1, ..., P$), up to first-order approximation:

$$\hat{R}_p = (I + \mathcal{E})SP_pS^T(I + \mathcal{E}^T) \approx SP_pS^T + \mathcal{E}SP_pS^T + SP_pS^TE^T, \hspace{1cm} (19)$$

where we have neglected terms which are quadratic in $\mathcal{E}$.

We show in Appendix A that under this assumption of nearly-separated sources, up to a first-order approximation (neglecting second and higher order terms in the elements of $\mathcal{E}$), we have:

$$E[\hat{r}_{k,\ell}] = \begin{cases} 
  h_k & k = \ell \\
  h_\ell \cdot \mathcal{E}[k, \ell] + h_k \cdot \mathcal{E}[\ell, k] & k \neq \ell
\end{cases}, \hspace{1cm} (20)$$

where $h_k \in \mathbb{R}^P$ is a vector whose elements are given by

$$h_k[p] = \text{Tr}\{P_pC_k\}, \hspace{0.5cm} p = 1, ..., P, \hspace{1cm} (21)$$

where $\text{Tr}\{\cdot\}$ denotes the trace operator.

In addition, we show in the Appendix that up to zero-order approximation (neglecting first and higher order terms in the elements of $\mathcal{E}$), the vectors $\hat{r}_{k,\ell}$ are uncorrelated with each other for all $k$ and $\ell$, namely for any quadruple $k, \ell, k', \ell'$ such that either $k \neq k'$ or $\ell \neq \ell'$ (or both), the vector $\hat{r}_{k,\ell}$ is uncorrelated with the vector $\hat{r}_{k',\ell'}$. The covariance matrix of $\hat{r}_{k,\ell}$ is denoted $K_{k,\ell} \in \mathbb{R}^{P \times P}$, and its elements are given (up to zero-order approximation) by

$$K_{k,\ell}[p, q] = \text{Tr}\{P_pC_\ell P_qC_k\}, \hspace{0.5cm} p, q \in [1, P] \hspace{1cm} (22)$$

for $k \neq \ell$. For $k = \ell$ the covariance $K_{k,k}$ involves fourth-order joint moments of $s_k$ (which can be expressed in terms of $C_k$ when $s$ is Gaussian), but the explicit
expression is of little interest here, because it is irrelevant to our subsequent derivations.

We shall now derive the approximate correlations-based ML estimate of the residual mixing-matrix \( I + E \) (namely of \( E \)) from \( \hat{q} \) and \( \hat{r} \). To summarize, our derivation will be based on the following approximating assumptions (to which we shall refer as the “model approximations”):

1. All elements of the new “front-end” data, namely the DIT and the ODIT vectors \( \hat{q} \) and \( \hat{r} \), are jointly Gaussian;
2. The mean of each vector-component \( \hat{r}_{k,\ell} \) of the DIT and ODIT vectors is given by the first-order approximation (20);
3. The joint covariance matrix of the DIT and ODIT vectors is given by the zero-order approximation (22), namely this matrix is block-diagonal with each \( P \times P \) block \( K_{k,\ell} \) accounting for the covariance of the respective \( P \times 1 \) vector-component \( \hat{r}_{k,\ell} \);
4. The vectors \( h_k \) and the covariance matrices \( K_{k,\ell} \) (for all \( k \leq \ell \)) are considered as additional unknown (but free) parameters, as a substitute to the much larger set of true “underlying” unknown parameters - namely the elements of the sources’ covariance matrices \( C_k \). This assumption essentially ignores the specific dependence of all \( h_k \) and \( K_{k,\ell} \) (a total of \( PK + \frac{1}{2}K(K+1) \cdot \frac{1}{2}P(P+1) \) free parameters) on all \( C_k \) (a total of \( K \cdot \frac{1}{2}N(N+1) \) free parameters). Since normally \( P, K, PK \ll N \), ignoring the inter-dependence between these sets of parameters is a reasonable approximation, since the set of possible values of all \( h_k \) and \( K_{k,\ell} \) is not really constrained by its relation to the far larger-dimensional covariance matrices \( C_k \).

Based on these model approximations, the joint probability distribution
function of $\hat{q}$ and $\hat{r}$ can be approximated as follows:

$$f_{\hat{q},\hat{r}}(\hat{q}, \hat{r}; \mathcal{E}, h_1, ..., h_K, \mathcal{K}_{1,1}, ..., \mathcal{K}_{K,K}) = f_{\hat{q}}(\hat{q}; h_1, ..., h_K, \mathcal{K}_{1,1}, ..., \mathcal{K}_{K,K}) \cdot f_{\hat{r}}(\hat{r}; \mathcal{E}, h_1, ..., h_K, \mathcal{K}_{1,2}, ..., \mathcal{K}_{K-1,K}),$$  \hspace{1cm} (23)

where we have exploited the fact that $\hat{q}$ and $\hat{r}$ are uncorrelated and therefore (being jointly Gaussian) independent, as well as the fact that the distribution of $\hat{q}$ does not depend on $\mathcal{E}$. Taking the log-likelihood of each term,

$$L_q(\hat{q}) \triangleq \log f_{\hat{q}}(\hat{q}; h_1, ..., h_K, \mathcal{K}_{1,1}, ..., \mathcal{K}_{K,K}) = -\frac{1}{2} \sum_{k=1}^{K} \left( \log |2\pi \mathcal{K}_{k,k}| + (\hat{r}_{k,k} - h_k)^T \mathcal{K}_{k,k}^{-1} (\hat{r}_{k,k} - h_k) \right)$$  \hspace{1cm} (24)

(where $|\cdot|$ denotes the determinant) and

$$L_r(\hat{r}) = \log f_{\hat{r}}(\hat{r}; \mathcal{E}, h_1, ..., h_K, \mathcal{K}_{1,2}, ..., \mathcal{K}_{K-1,K}) = -\frac{1}{2} \sum_{k,\ell=1}^{K} \left( \log |2\pi \mathcal{K}_{k,\ell}| + (\hat{r}_{k,\ell} - h_\ell \cdot \mathcal{E}[k,\ell] - h_k \cdot \mathcal{E}[\ell,k])^T \mathcal{K}_{k,\ell}^{-1} (\hat{r}_{k,\ell} - h_\ell \cdot \mathcal{E}[k,\ell] - h_k \cdot \mathcal{E}[\ell,k]) \right).$$  \hspace{1cm} (25)

We now reinstate the multiple-snapshots framework, and exploiting the statistical independence between snapshots, our ultimate goal is to maximize, w.r.t. all of the parameters $\mathcal{E}, h_1, ..., h_K, \mathcal{K}_{1,1}, \mathcal{K}_{1,2}, ..., \mathcal{K}_{K,K}$ the sum of likelihoods

$$L \triangleq \sum_{m=1}^{M} \left( L_q(\hat{q}^{(m)}) + L_r(\hat{r}^{(m)}) \right),$$  \hspace{1cm} (26)

where $\hat{q}^{(m)}$ and $\hat{r}^{(m)}$ denote the DIT and ODIT vectors (resp.) taken from the $m$-th snapshot.

Due to the nonlinearity of the overall likelihood in the entire set of parameters, joint maximization w.r.t. all of the parameters together would be rather cumbersome (and partly unnecessary, as we shall explain immediately). We therefore apply an approximate maximization procedure taking the following steps (which we term the “maximization approximations”):
1. First, we set $\mathbf{E}$ to its zero-order approximate value of all zeros, $\mathbf{E} = \mathbf{0}$;

2. Consequently, the dependence on $h_1, ..., h_K$ is now only reflected in

$$
\sum_{m=1}^{M} L_q(q^{(m)}) = C - \frac{1}{2} \sum_{k=1}^{K} \sum_{m=1}^{M} (\hat{r}_{k,k}^{(m)} - h_k)^T \mathbf{K}_{k,k}^{-1} (\hat{r}_{k,k}^{(m)} - h_k), \quad (27)
$$

(where $C$ is an irrelevant constant), such that maximization with respect to these vectors is evidently attained (assuming $M > P$) by setting each $h_k$ to the snapshots-average

$$
\hat{h}_k = \frac{1}{M} \sum_{m=1}^{M} \hat{r}_{k,k}^{(m)}, \quad (28)
$$

regardless of the values of the covariance matrices $\mathbf{K}_{k,k}$ (since they do not depend on $m$);

3. Moreover, the covariance matrices of type $\mathbf{K}_{k,k}$ are no longer relevant for maximization of $L$ w.r.t. $\mathbf{E}$ and are therefore of no interest;

4. In addition, maximization w.r.t. the covariance matrices of type $\mathbf{K}_{k,\ell}$ merely requires the maximization of

$$
\sum_{m=1}^{M} L_r(\hat{r}^{(m)}) = -\frac{1}{2} \sum_{k,\ell=1}^{K} \sum_{m=1}^{M} \left( \log[2\pi |\mathbf{K}_{k,\ell}|] + (\hat{r}_{k,\ell}^{(m)})^T \mathbf{K}_{k,\ell}^{-1} (\hat{r}_{k,\ell}^{(m)}) \right), \quad (29)
$$

which (since this expression can be decoupled into distinct $(k, \ell)$ terms) is evidently attained by the well-known ML covariance estimates

$$
\hat{\mathbf{K}}_{k,\ell} = \frac{1}{M} \sum_{m=1}^{M} (\hat{r}_{k,\ell}^{(m)}) (\hat{r}_{k,\ell}^{(m)})^T; \quad (30)
$$

5. The final (and most important) step is the maximization of $L$ w.r.t. $\mathbf{E}$, once the estimates of all $h_k$ and $\mathbf{K}_{k,\ell}$ have been obtained. Evidently, the only term relevant for this maximization is now

$$
\sum_{m=1}^{M} L_r(\hat{r}^{(m)}) = C' - \frac{1}{2} \sum_{k,\ell=1}^{K} \sum_{m=1}^{M} (\hat{r}_{k,\ell}^{(m)} - \hat{h}_k : \mathbf{E}[k, \ell] - \hat{h}_k : \mathbf{E}[\ell, k])^T \hat{\mathbf{K}}_{k,\ell}^{-1} (\hat{r}_{k,\ell}^{(m)} - \hat{h}_k : \mathbf{E}[k, \ell] - \hat{h}_k : \mathbf{E}[\ell, k])
$$

(31)
(where $C'$ is an irrelevant constant). This expression can be decoupled into distinct $(k, \ell)$ terms, since both $\mathcal{E}[k, \ell]$ and $\mathcal{E}[\ell, k]$ are only present in exactly one combination of $(k, \ell)$ ($k < \ell$), so for each $(k, \ell)$ combination we only need to minimize the inner sum in (31). Being quadratic in $\mathcal{E}[k, \ell]$ and $\mathcal{E}[\ell, k]$, the minimization is easily identified as a Weighted LS (WLS) problem (with the inverse covariance matrices $\hat{\mathcal{K}}_{k,\ell}^{-1}$ as weights), to which the well-known solution is obtained (for each $(k, \ell)$, $k < \ell$ pair) as

$$
\begin{bmatrix}
\mathcal{E}[k, \ell] \\
\mathcal{E}[\ell, k]
\end{bmatrix} = 
\begin{bmatrix}
\hat{h}_T \hat{K}_{k,\ell}^{-1} \hat{h}_\ell & \hat{h}_T \hat{K}_{k,\ell}^{-1} \hat{h}_k \\
\hat{h}_T \hat{K}_{k,\ell}^{-1} \hat{h}_\ell & \hat{h}_T \hat{K}_{k,\ell}^{-1} \hat{h}_k
\end{bmatrix}^{-1}
\begin{bmatrix}
\hat{h}_T \hat{K}_{k,\ell}^{-1} \hat{r}_k,\ell \\
\hat{h}_T \hat{K}_{k,\ell}^{-1} \hat{r}_k,\ell
\end{bmatrix},
$$

(32)

where

$$
\hat{r}_k,\ell \triangleq \frac{1}{M} \sum_{m=1}^{M} \hat{r}_{k,\ell}^{(m)}.
$$

(33)

are the components of the “average ODIT vector”, averaged over all snapshots.

In order to proceed with the maximization, it is possible to repeat the above stages with the new estimated elements of $\mathcal{E}$ replacing the zero-order approximation, but then the minimization w.r.t. all $\hat{h}_k$ would require more involved expressions than on Step 2 above, because the $\hat{h}_k$ terms would then be present both in the $L_q(\cdot)$ and $L_r(\cdot)$ terms. Alternatively, it is possible to use the estimated $\mathcal{E}$ so as to further separate the sources (refining the initial separation), thereby obtaining a new approximate ML problem, to which the above approximate maximization procedure can be applied with the all-zeros initialization again. We note, however, that if the initial separation is sufficiently successful, such further iterations would not change the final estimate of the mixing (or demixing) matrix, namely, subsequent iterations would usually result in extremely small values in $\mathcal{E}$.

In fact, while the above procedure (together with the associated approximations) was developed from the perspective of CBML estimation, it can also be regarded as a weighted AJD process applied to the (snapshots average of) the generalized correlation matrices $\hat{R}_1, ..., \hat{R}_P$. Indeed, such a weighted AJD
process also requires the estimation of the covariance matrices \( \mathbf{K}_{k,\ell} \) from the multiple snapshots. Once these covariance matrices are estimated, the use of their inverses for the weighted AJD problem conveniently lends itself to the recently-proposed WEDGE algorithm \([11]\) for AJD with such block-diagonal weighting.

The possibility to estimate these covariance matrices directly from the observed data, without any prior information on the sources’ temporal covariance matrices \( \mathbf{C}_k \), enables to approach the optimal performance (although usually still does not enable to reach that performance) in cases where alternative model-based optimal approaches (such as WASOBI \([12, 11]\)) suffer from model mismatch. We demonstrate this ability in Section 4.

A convenient by-product of the WLS formulation is an asymptotic expression for the expected performance. Assuming that the estimated \( \hat{\mathbf{h}}_k, \hat{\mathbf{h}}_\ell \) and \( \hat{\mathbf{K}}_{k,\ell} \) are close (asymptotically) to their respective true values (prescribed in \((20), (22))\), the error-covariance of the WLS estimate \((32)\) of \( \hat{\mathbf{E}}[k, \ell] \) and \( \hat{\mathbf{E}}[\ell, k] \) is given by

\[
\text{cov} \begin{pmatrix} \hat{\mathbf{E}}[k, \ell] \\ \hat{\mathbf{E}}[\ell, k] \end{pmatrix} = \frac{1}{M} \begin{bmatrix} \hat{\mathbf{h}}_k^T \hat{\mathbf{K}}_{k,\ell}^{-1} \hat{\mathbf{h}}_\ell & \hat{\mathbf{h}}_k^T \hat{\mathbf{K}}_{k,\ell}^{-1} \hat{\mathbf{h}}_k \\ \hat{\mathbf{h}}_k^T \hat{\mathbf{K}}_{k,\ell}^{-1} \hat{\mathbf{h}}_\ell & \hat{\mathbf{h}}_\ell^T \hat{\mathbf{K}}_{k,\ell}^{-1} \hat{\mathbf{h}}_\ell \end{bmatrix}^{-1}, \tag{34} \]

so that

\[
E \left[ \hat{\mathbf{E}}^2[k, \ell] \right] = \frac{1}{M} \frac{\hat{\mathbf{h}}_k^T \hat{\mathbf{K}}_{k,\ell}^{-1} \hat{\mathbf{h}}_k}{(\hat{\mathbf{h}}_k^T \hat{\mathbf{K}}_{k,\ell}^{-1} \hat{\mathbf{h}}_k)(\hat{\mathbf{h}}_\ell^T \hat{\mathbf{K}}_{k,\ell}^{-1} \hat{\mathbf{h}}_\ell) - (\hat{\mathbf{h}}_k^T \hat{\mathbf{K}}_{k,\ell}^{-1} \hat{\mathbf{h}}_k)^2}. \tag{35} \]

This value is the \((k, \ell)\)-th element of the ISR matrix\(^3\) when the \(k\)-th and \(\ell\)-th sources have the same energy. When this is not the case, this expression has to be multiplied by \(\text{Tr}\{\mathbf{C}_\ell\}/\text{Tr}\{\mathbf{C}_k\}\) to yield the respective element of the ISR matrix (see, e.g., \([15]\)).

\(^3\)The ISR matrix contains, in its \((k, \ell)\)-th element, the mean ratio between the residual energy of the \(\ell\)-th source in the reconstruction of the \(k\)-th source, and the energy of the latter. Under a small-errors assumption, this mean ratio is given by \((35)\) when the sources have equal energies. Otherwise, this expression should be normalized by the energy ratios of the \(\ell\)-th and \(k\)-th sources.
We reiterate, however, that this expression is only asymptotic, since it ignores the effect of the estimation errors in $\hat{h}_k$, $\hat{h}_\ell$ and $\hat{K}_{k,\ell}$, which might be significant under non-asymptotic conditions. In fact, it can be shown that this expression coincides with the Cramér-Rao bound on estimation of $\mathcal{E}$ from the SGC matrices under our “model approximation” assumptions. This bound should not be confused with the Cramér-Rao bound on estimation of $\mathcal{E}$ from the original observations, which would always be equal or lower (sometimes much lower). While both can only be attained asymptotically (in $M$), the former would be approached much sooner (namely for a smaller $M$), since its model involves considerably fewer unknown parameters.

We conclude this section with a brief summary of the CBML estimation process in the Table below.
Apply some initial (consistent) separation algorithm such as SOBI [1] or WASOBI [11] (preferably, use correlation estimates averaged over all available snapshots); denote the estimated separation matrix as $\hat{B}_0$ and the estimated separated sources in each snapshot as: 

$$\hat{S}^{(m)} = \hat{B}_0 X^{(m)}, \ m \in [1, M].$$

1. Obtain the $P$ SGC matrices in each snapshot: 

$$\hat{R}_{p}^{(m)} = \hat{S}^{(m)} P_p (\hat{S}^{(m)})^T \in \mathbb{R}^{K \times K}, \ p \in [1, P], \ m \in [1, M].$$

2. Obtain the vector-components of the DIT and ODIT vectors at each snapshot ($m \in [1, M]$) as: 

$$\hat{r}_{k, \ell}^{(m)} = [R_1^{(m)}(k, \ell) \cdots R_p^{(m)}(k, \ell)]^T \in \mathbb{R}^P, \ k \leq \ell \in [1, K].$$

3. Estimate the $\frac{1}{2}K(K - 1)$ covariance matrices of the vector-components of the ODIT vector as follows: 

$$\hat{K}_{k, \ell} = \frac{1}{M} \sum_{m=1}^{M} (\hat{r}_{k, \ell}^{(m)}) (\hat{r}_{k, \ell}^{(m)})^T \in \mathbb{R}^{P \times P}, \ k < \ell \in [1, K].$$

4. Obtain the mean vectors $\hat{h}_k \triangleq \frac{1}{M} \sum_{m=1}^{M} \hat{r}_{k, k}^{(m)} \in \mathbb{R}^P, \ k \in [1, K]$, and $\hat{v}_{k, \ell} \triangleq \frac{1}{M} \sum_{m=1}^{M} \hat{r}_{k, \ell}^{(m)} \in \mathbb{R}^P, \ k < \ell \in [1, K]$. 

5. Using all $\hat{h}_k$, $\hat{v}_{k, \ell}$ and $\hat{K}_{k, \ell}$, solve the $\frac{1}{2}K(K - 1)$ sets of $2 \times 2$ equations (32) for all $k < \ell \in [1, K]$, obtaining elements $\hat{E}[k, \ell]$ of $\hat{E}$ for all $k \neq \ell$; Set all diagonal elements $\hat{E}[k, k]$ to zero. 

6. The estimated residual demixing matrix is given by $\hat{B}_1 = I + \hat{E}$. The overall estimated demixing matrix is given by $\hat{B} = \hat{B}_1 \hat{B}_0$. 

Outputs: The estimated demixing matrix $\hat{B}$. The separated sources (in each snapshot) can be obtained either by applying $\hat{B}$ to all $X^{(m)}$ or by applying $\hat{B}_1$ to all $\hat{S}^{(m)}$.

If desired, the process can be repeated by regarding the attained separation at Step 7 as the initial separation and returning to Step 1 (setting $\hat{B}_0 = \hat{B}$). Note that when so doing, it is not actually necessary to re-compute the separated

<table>
<thead>
<tr>
<th>CBML</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inputs:</strong> $M$ snapshots of the mixtures $X^{(m)} \in \mathbb{R}^{K \times N}, \ m \in [1, M]$; $P$ association matrices $P_1, \ldots, P_p \in \mathbb{R}^{N \times N}$.</td>
</tr>
<tr>
<td><strong>Outputs:</strong> The estimated demixing matrix $\hat{B}$. The separated sources (in each snapshot) can be obtained either by applying $\hat{B}$ to all $X^{(m)}$ or by applying $\hat{B}_1$ to all $\hat{S}^{(m)}$.</td>
</tr>
</tbody>
</table>
sources (Step 1) in each iteration, since the SGC matrices (for Step 2) can be updated directly by multiplication with \( \hat{B} \) on the left and with \( \hat{B}^T \) on the right. The very same iterative update approach is taken in the WEDGE algorithm for weighted AJD [11].

4. Simulation results

We demonstrate the potential advantage of CBML over DBML using a mixture of \( K = 4 \) independent nonstationary time-varying Auto-Regressive Moving-Average (ARMA) processes with drifting poles and zeros as follows.

Each source signal \( s_k[n] \) of length \( N = 100 \) is generated using the following generation model

\[
s_k[n] = -6 \sum_{\ell=1}^{6} a_k[n, \ell] s_k[n - \ell] + w_k[n] + 4 \sum_{\ell=1}^{4} b_k[n, \ell] w_k[n - \ell]
\]

for \( n = 1, \ldots, 100 \), with zero initial conditions (namely, zero values of \( s_k[n] \) for \( n \leq 0 \)), where \( w_k[n] \) are i.i.d. zero-mean unit-variance Gaussian “driving-noise” sequences. Thus, all the source signals are Gaussian time-varying ARMA(6,4) processes whose time-varying AR parameters \( \{a_k[n,1], \ldots, a_k[n,6]\} \) and MA parameters \( \{b_k[n,1], \ldots, b_k[n,4]\} \) (for \( 1 \leq k \leq K \) and \( 1 \leq n \leq N \)) were set so as to satisfy

\[
1 + 6 \sum_{\ell=1}^{6} a_k[n, \ell] z^{-\ell} = 6 \prod_{\ell=1}^{6} (1 - \xi_k[n, \ell] z^{-1})
\]

\[
1 + 4 \sum_{\ell=1}^{4} b_k[n, \ell] z^{-\ell} = 4 \prod_{\ell=1}^{4} (1 - \zeta_k[n, \ell] z^{-1}),
\]

with the respective time-varying poles \( \{\xi_k[n,1], \ldots, \xi_k[n,6]\} \) and time-varying zeros \( \{\zeta_k[n,1], \ldots, \zeta_k[n,4]\} \) drifting along circles (in Z-plane) of radii 0.4 (for all poles and zeros), such that their phases drift linearly with \( n \) from an initial phase \( \varphi_i \) at \( n = 1 \) to a final phase \( \varphi_f \) at \( n = N \). The initial and final phases of the poles and zeros are summarized in Table 1 below.

Since all signals (and coefficients) are real-valued, the (inverted) phases of half of the poles and zeros are omitted from the table, as all complex-valued
Table 1: Initial and final phases (in degrees) of the time-varying poles (note that $s_2[n]$ has constant poles and zeros, but is still nonstationary, due to initial conditions.

<table>
<thead>
<tr>
<th>source id</th>
<th>pole 1</th>
<th>pole 2</th>
<th>pole 3</th>
<th>zero 1</th>
<th>zero 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 1$</td>
<td>$\varphi_i$</td>
<td>$\varphi_f$</td>
<td>$\varphi_i$</td>
<td>$\varphi_f$</td>
<td>$\varphi_i$</td>
</tr>
<tr>
<td>$k = 2$</td>
<td>150</td>
<td>10</td>
<td>160</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>150</td>
<td>10</td>
<td>160</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>$k = 4$</td>
<td>150</td>
<td>10</td>
<td>160</td>
<td>20</td>
<td>30</td>
</tr>
</tbody>
</table>

poles and zeros must appear in complex-conjugate pairs. Figure 2 illustrates a typical realization of the four sources.

Figure 2: (appended at the end) Sample realizations of the four non-stationary sources.

We mixed the four sources with a random $4 \times 4$ mixing-matrix $A$ with i.i.d. standard Gaussian elements, redrawn for each trial (but held fixed for all snapshots of the same trial). The separation results are presented in Figure 3 in terms of the individual ISR values $\text{ISR}[1, 2]$, $\text{ISR}[1, 3]$ and $\text{ISR}[1, 4]$ vs. the number $M$ of snapshots per trial, reflecting averaged performance over 1000 independent trials. The empirical $\text{ISR}[k, \ell]$ element is obtained as follows. Let $A_t$ and $\hat{B}_t$ denote the true mixing matrix and estimated demixing matrix (resp.) at the $t$-th trial, with the residual permutation and scale ambiguities in $\hat{B}_t$ resolved based on knowledge of the true sources. We define $T_t \triangleq \hat{B}_t A_t$, the overall mixing-demixing matrix for the $t$-th trial. The $(k, \ell)$-th element of the ISR matrix is then obtained as

$$\text{ISR}[k, \ell] = \left( \frac{1}{1000} \sum_{t=1}^{1000} T_t^2[k, \ell] \right) \cdot \frac{\text{Tr}\{C_\ell\}}{\text{Tr}\{C_k\}}, \quad k \neq \ell \in [1, 4] \quad (38)$$

which represents the relative residual energy of the $\ell$-th source in the reconstruction of the $k$-th sources. In addition, we present in Figure 4 the mean total ISR value, which is the total ISR value per source, averaged among all sources, $\frac{1}{K} \sum_{k \neq \ell} \text{ISR}[k, \ell]$. 

22
In both figures we compare the following:\footnote{For Figure 4 we show the mean total values of all these quantities.}

1. The induced Cramér-Rao bound on the ISR, derived in [15] and given by
   \[
   \text{CRB}_k[\ell] = \frac{1}{N} \cdot \frac{\phi_{\ell,k}}{\phi_{\ell,k} + 1} \cdot \frac{\text{Tr}\{C_\ell\}}{\text{Tr}\{C_k\}} \quad k \neq \ell \in [1, K],
   \]  
   where \( \phi_{k,\ell} \) is defined as \( \phi_{k,\ell} \triangleq \frac{1}{N} \text{Tr}\{C_{\ell}^{-1}C_k\} \);

2. The semi-blind ML estimate (obtained using the true covariance matrices of the sources);

3. The “full” DBML estimate, based on individual ML estimates (sample-estimates) of the unknown \( N \times N \) covariance matrices \( C_k \) from the multiple snapshots (after initial separation). Note that these results are only available for \( M \geq N = 100 \), since for \( M < N \) the sample-covariances are not invertible (and are not ML estimates), and therefore cannot be substituted into the “target matrices” (8) for estimation of the demixing matrix;

4. The proposed CBML estimate using the following seven SGC matrices (\( P = 7 \)): lagged sample-correlation matrices at lags 0 to 5, and an additional matrix of sample zero-lag correlations, estimated only over the first half (50 samples) of the observation interval, applying the most “natural” division into two blocks in an attempt to capture the nonstationarity; The CBML algorithm was applied as summarized above, without adding additional iterations;

5. The predicted ISR of the CBML estimate, using (35) (with energy normalization);

6. For reference, we also present results of the fully-blind AR-WASOBI [11], which is (asymptotically) optimal for stationary AR sources, but obviously not optimal with our non-stationary time-varying ARMA sources. The algorithm was applied with the same number of \( P = 7 \) lagged sample-correlation matrices at lags 0 to 6.
7. For additional reference, we present results of applying an algorithm based on selected time-frequency distribution matrices, proposed in [8]. While, as mentioned above, time-frequency distribution matrices can be regarded as particular choices of the $P_p$ matrices, the joint diagonalization approach in [8] is different, as it uses Pham’s Gaussian Mutual Information criterion. The algorithm was applied with a Hamming temporal window of length 50, a Parzen window of length 25, and using six time-frequency points composed of all six combinations of $t \in \{25, 75\}$ with $\omega \in \{0, \frac{\pi}{3}, \frac{2\pi}{3}\}$ (see [8] for more implementation details).

Evidently, the semi-blind full-ML estimate is very close to the bound for all ISRs, as could be expected. However, among the fully-blind estimators, CBML significantly outperforms the others in $ISR_{1,2}$ and in $ISR_{1,4}$ for the lower values of $M$. In $ISR_{1,3}$ CBML is only slightly better than WASOBI, and is slightly outperformed by the Time-Frequency method of [8]. This indicates that for separating $s_3[n]$ from $s_1[n]$, the ordinary lagged correlations (for WASOBI), or our selected time-frequency distribution matrices (for [8]), already attain good separation on one hand (note that the respective $ISR_{1,3}$ values are significantly smaller than the other ISR values for each $M$), but are still far from optimal on the other hand. Generally similar behavior is observed in the mean total ISR, but with the method of [8] performing much closer to CBML, mainly due to the fact that the selected time-frequency points perform well for the worse ISRs, which dominate the total ISR (the worse ISRs happen to be $ISR_{2,4}$ and $ISR_{4,2}$ in our setup; they are about $10dB$ worse than the next worse ISRs).

It is also observed from the slope of the DBML estimate, that DBML would eventually approach the bound as well, but this would only occur for signifi-
cantly higher values of $M$ for some of the ISR elements, and for much larger (often unfeasible) values of $M$ for the mean total ISR. Thus, for a very large number of snapshots $M$ (orders higher than $10^5$ in this case), the consistent ML estimates of the $C_k$ in DBML would make the DBML estimate coincide with the semi-blind ML estimate, which attains the bound - while CBML and AR-WASOBI would maintain their sub-optimality even at infinitely large values of $M$. Nevertheless, under our more realistic sub-asymptotic conditions, DBML is considerably farther from optimality than CBML. Note also that the empirical CBML performance is seen to approach (at least asymptotically) its analytically-predicted performance (which was derived under the small-errors assumption).

5. Conclusion

Asymptotically-optimal separation (in the sense of minimum ISR) of Gaussian sources with arbitrary temporal covariance structures requires either perfect knowledge, or consistent estimation of the temporal covariance matrices. While such estimation is theoretically possible if several snapshots are available, the required number of snapshots for reliable ML estimation might be huge (of orders much larger than the squared observation length).

In this work we proposed an alternative approach for exploiting a more realistic number of snapshots in a “partial” ML approach, termed CBML, which considers the observations’ spatial sample-generalized-correlation matrices (following some initial separation stage) as the “front-end” data for ML estimation. These matrices are prescribed by a chosen set of “association matrices”, and can be composed, e.g., of lagged sample-correlation matrices at different lags, as sample-correlation matrices taken over sub-intervals of the observation interval, or as arbitrarily many other forms. Once such a set is chosen, the resulting CBML estimate was shown to require the estimation of much smaller covariance matrices, whose inverses are essentially used in pairwise WLS estimation of the elements of the residual demixing matrix. These estimates are equivalent to those obtained in weighted joint diagonalization of the same SGC matrices.
As an important by-product, we provided analytic expressions for the expected ISR performance of CBML, derived under a small-errors assumption.

Possible applications for using the proposed approach are context in which triggered multiple-snapshots mixtures can be available, such as in functional Magnetic Resonance Imaging (fMRI), where measurements are usually triggered to an external stimulus, or in similar experiments in electroencephalography (EEG). Most of the signals involved in such experiments are usually nonstationary, but with no prescribed structure.

It should be noted, that when asymptotics are addressed in the context of our multiple-snapshots scenario, there are several possible interpretations of the term. Our claim of asymptotic optimality of the CBML (with respect to the selected set of SGC matrices) refer to the case where \( N \) is fixed and \( M \) tends to infinity.

On the other hand, if \( M \) is fixed and \( N \) tends to infinity, then the situation is essentially similar to the single-snapshot scenario (but with possibly faster convergence if \( M > 1 \)). Asymptotic assumptions may hold if the sources are “persistently excited” in some sense (e.g., being stationary, or nearly stationary, or cyclo-stationary, etc.) and the association matrices \( P_p \) are properly selected (satisfying some richness and condition-number conditions). Similar restrictions apply when both \( M \) and \( N \) tend to infinity (possibly with a fixed ratio).

A remaining open question in this context is how to choose the SGC matrices, or the association matrices \( P_1, \ldots, P_P \) (and their number \( P \)). As already mentioned, if the sources are known to be Gaussian AR processes, then the optimal selection of association matrices would be the Toeplitz matrices yielding the lagged correlations up to the maximal AR order of the sources, since such matrices form (asymptotically) a sufficient statistic with respect to the data (see [5, 11]). In other cases, even if the sources are not AR processes, but are at least “somewhat” stationary, it appears that lagged correlations would still make very reasonable SGC matrices; however, if any localized information about the sources is available, then the addition of respectively localized SGC matrices (such as in our simulation) would usually be helpful as well. It is also
evident from our simulation results in comparing to the time-frequency method of [8], that a selected set of time-frequency points (which merely implies a particular association matrix) can also be quite informative. An important point to keep in mind is that under asymptotic conditions in $M$, the SGC matrices are optimally weighted for the estimation, and therefore adding more and more SBC matrices would not degrade (and would usually improve) the performance. But obviously, this is not necessarily the case when $M$ is not sufficiently large, so wise selection of the association matrices is still an important issue in such cases, for which we do not (yet?) have a proposed strategy in general.

Although definitely sub-optimal under asymptotic conditions (in the number of snapshots), our proposed method was demonstrated to offer significant improvement over the asymptotically-optimal “full” ML estimate when the number of available snapshots is limited.

Appendix A. Mean and covariance of the DIT and ODIT vectors
We begin with the first-order approximation (19), repeated for convenience,

$$\hat{R}_p = S P_p S^T + \mathcal{E} S P_p S^T + S P_p S^T \mathcal{E}^T,$$  
$p = 1, ..., P.$  

(A.1)

Let us define the matrix $D_p \triangleq E[S P_p S^T]$, such that taking the mean of (A.1) we have $E[\hat{R}_p] = D_p + \mathcal{E} D_p + D_p \mathcal{E}^T$ (for $p = 1, ..., P$). The $(k, \ell)$-th element of $D_p$ is given by

$$D_p[k, \ell] = E[s_k^T P_p s_\ell] = \text{Tr}\{P_p E[s_\ell s_k^T]\} = \begin{cases} 
0 & k \neq \ell \\
\text{Tr}\{P_p C_k\} & k = \ell 
\end{cases}.$$  

(A.2)

Recalling that the diagonal elements of $\mathcal{E}$ are assumed to be all zeros (since they can be absorbed by the inherent scaling ambiguity), we obtain the following explicit expressions for the elements of the mean matrix:

$$E[R_p[k, \ell]] = \begin{cases} 
\mathcal{E}[k, \ell] D_p[\ell, \ell] + D_p[k, k] \mathcal{E}[\ell, \ell] & k \neq \ell \\
D_p[k, k] & k = \ell 
\end{cases}.$$  

(A.3)

27
Recalling (from (15)) the structure of \( \hat{r}_{k,\ell} = \left[ \hat{R}_1[k,\ell] \ldots \hat{R}_P[k,\ell] \right]^T \in \mathbb{R}^p \), we identify the elements of the mean vectors by substituting (A.3) and (A.2)

\[
E[\hat{r}_{k,\ell}] = \begin{cases} h_k & k = \ell \\ h_\ell \cdot \mathcal{E}[k,\ell] + h_k \cdot \mathcal{E}[\ell,k] & k \neq \ell \end{cases},
\]

(A.4)

where the elements of \( h_k \in \mathbb{R}^p \) are \( h_k[p] = \text{Tr}\{P_p C_k\} \).

Moving on to the covariance expressions, we employ a zero-order approximation, assuming that \( \mathcal{E} = 0 \), such that \( \hat{R}_p = S P_p S^T \). Seeking second joint moments we have, for \( p, p' \in [1, P] \) and \( k, \ell, k', \ell' \in [1, K] \):

\[
E[\hat{R}_p[k,\ell] \hat{R}_{p'}[k',\ell']] = E[sp_k^T P_p s_{k'}^T P_{p'} s_{\ell'}] = \text{Tr}\{P_p s_{k}^T P_{p'} s_{k'}^T \text{Tr}\{P_p C_{k}\}} \text{Tr}\{P_{p'} C_{k'}\};
\]

(A.5)

Obviously, if any one of the indices \( k, \ell, k', \ell' \) is different from all the others, then by rearranging the terms on the right-hand side in (A.5) it is possible to isolate the distinct index, and then, due to the sources’ independence, the mean of the product equals the product of the two means - which is zero. This means that the covariance between any two vectors \( \hat{r}_{k,\ell} \) and \( \hat{r}_{k',\ell'} \) vanishes for every combination of \( k, \ell, k', \ell' \) except for combinations admitting the following partitions into couples:

For \( (k = \ell) \neq (k' = \ell') \):

\[
E[\hat{R}_p[k,k] \hat{R}_{p'}[k',k']] = E[sp_k^T P_p s_{k'}^T P_{p'} s_{k'}] = \text{Tr}\{P_p C_k\} \text{Tr}\{P_{p'} C_{k'}\};
\]

Such cases refer to joint second moments between different elements \( \hat{r}_{k,k} \) and \( \hat{r}_{k',k'} \) of the DIT vector. Since this joint second moment equals the product of the respective means (under the zero-order approximation), we deduce that \( \hat{r}_{k,k} \) and \( \hat{r}_{k',k'} \) are uncorrelated.

For \( (k = k') \neq (\ell = \ell') \):

\[
E[\hat{R}_p[k,\ell] \hat{R}_{p'}[k,\ell]] = E[sp_k^T P_p s_{k}^T P_{p'} s_{\ell}] = E[sp_k^T P_p s_{k}^T P_{p'} s_{\ell}] = \text{Tr}\{E[|P_p s_k s_{k}^T P_{p'} s_{\ell}]\} = \text{Tr}\{E[|P_p s_k s_{k}^T P_{p'} s_{\ell}]\} = \text{Tr}\{P_p C_k P_{p'} C_{\ell}\};
\]

28
Such cases refer to joint second moments within elements $\hat{r}_{k,\ell}$ of the ODIT vector, which (under the zero-order approximation) all have zero mean - thereby defining the $(p, p')$-th elements of $K_{k,\ell}$, the covariance of $\hat{r}_{k,\ell}$.

For $k = k' = \ell = \ell'$:

$$E[\hat{R}_p[k, k] \hat{R}_{p'}[k, k]] = E[s_k^T P_p s_k s_k^T P_{p'} s_k].$$

Such cases refer to joint second moments within elements $\hat{r}_{k,k}$ of the DIT vector. These terms generally depend on the fourth-order joint moments of the respective sources’ distribution. Since the explicit expression is of no particular relevance to our CBML estimation, we shall not pursue the specific expression for the resulting covariance matrices of $\hat{r}_{k,k}$.

Note that we omitted the case $(k = \ell') \neq (k' = \ell)$ since the DIT and ODIT vectors only include combinations with $k \leq \ell$ and $k' \leq \ell'$, which are inconsistent with $(k = \ell') \neq (k' = \ell)$. In other words, we ignore the obvious relation between $\hat{r}_{k,\ell}$ and $\hat{r}_{\ell,k}$ (which are identical due to the symmetry of the SGC matrices) by excluding the latter for $k \leq \ell$.

References


Figure 2
Figure 3

ISR[1,2]

ISR[1,3]

ISR[1,4]

- induced CRLB
- semi-blind ML
- DBML
- CBML
- predicted CBML
- fully-blind WASOBI
- Pham’s T-F
Figure 4