Joint Independent Subspace Analysis Using Second-Order Statistics

Dana Lahat, Christian Jutten

Joint Independent Subspace Analysis Using Second-Order Statistics

Dana Lahat and Christian Jutten, *Fellow, IEEE*

***Abstract*—This paper deals with a novel generalization of classical blind source separation (BSS) in two directions. First, relaxing the constraint that the latent sources must be statistically independent. This generalization is well-known and sometimes termed independent subspace analysis (ISA). Second, jointly analysing several ISA problems, where the link is due to statistical dependence among corresponding sources in different mixtures. When the data are one-dimensional, i.e., multiple classical BSS problems, this model, known as independent vector analysis (IVA), has already been studied. In this work, we combine IVA with ISA and term this new model joint independent subspace analysis (JISA). We provide full performance analysis of JISA, including closed-form expressions for minimal mean square error (MSE), Fisher information and Cramér-Rao lower bound, in the separation of Gaussian data. The derived MSE applies also for non-Gaussian data, when only second-order statistics are used. We generalize previously-known results on IVA, including its ability to uniquely resolve static mixtures of real Gaussian stationary data, and having the same arbitrary permutation at all mixtures. Numerical experiments validate our theoretical results, and show the gain with respect to two competing approaches that either use a finer block partition or a different norm.**

***Index Terms*—Blind source separation, independent subspace analysis, coupled factorizations, data fusion, multiset, performance analysis.**

I. Introduction

In this work, we present a model inspired by two extensions to blind source separation (BSS) that until recently have been dealt with only separately: (1) relaxing the constraint that latent sources within a set of measurements must be statistically independent, often termed independent subspace analysis (ISA) or multidimensional independent component analysis (MICA) [1]–[3], and (2) solving several BSS problems simultaneously by exploiting statistical dependencies among latent sources across sets of measurements, a model known as independent vector analysis (IVA) or joint BSS (JBSS) [4], [5]. The new model, termed joint independent subspace analysis (JISA) [6], is a generalization of JBSS to multidimensional components.

The concept of ISA was first introduced in [1, Section 8], as the separation of several statistically independent random vectors. The idea that natural sources may be represented by multidimensional components such that only their corresponding subspaces have to be separated was first proposed in [2], who demonstrated it on fetal electrocardiography (ECG) recordings using an algebraic approach to independent component analysis (ICA). An elaborate geometric framework to the perspective of *multidimensional* ICA, whose focus is on vector-valued components whose representation is based on unambiguous projections on the sources' respective subspaces, was presented in [3]. A prevalent approach for ISA consists of using ICA-based algorithms followed by a clustering step [7]–[10]. Algorithms that exploit the true multidimensional nature of the data can be found, for example, in [11]–[18]. A theoretical analysis of the advantage, in terms of component estimation error, of using the true multidimensional model over the more prevalent two-step approach of BSS followed by a clustering step, is given in [19] for real Gaussian piecewise-stationary data. Identifiability and uniqueness of decompositions into invariant subspaces of dimensions larger than one are discussed in [20]–[24].

Multidimensional components may occur due to various complex relations and processes within the underlying phenomena that generate the data. A first example is neurological activity observed by functional magnetic resonance imaging (fMRI) [25] and electroencephalography (EEG) [26]. In convolutive mixtures, subspaces may represent channel effects (e.g., [27]). In audio and speech enhancement, subspaces can be used to model separate conversations, that is, disjoint groups of speakers [28]. Other types of phenomena that generate multidimensional components include astrophysical processes [29], fetal ECG [2] and natural images [11]. For such data, a one-dimensional model is often just an approximation. In the above-mentioned examples, the dimension of a dependent group may not always reflect the number of its underlying physical elements. Therefore, there is not always a physically meaningful interpretation to further separating the multidimensional components into single-dimensional elements. In this paper, we focus on separating subspaces that represent statistically independent multivariate components. Further decomposition, within a dependent group, if admissible by the application, is beyond the scope of this work.

One of the earliest frameworks to simultaneously analyse several data sets through statistical links between their latent parameters is canonical correlation analysis (CCA) [30]. The idea to simultaneously solve several ICA problems by exploiting higher-order statistical *dependence* among latent sources *across sets of measurements* was introduced by Kim et al. [4], and termed IVA. The method has been shown to be able to resolve the permutation ambiguity that is inherent to classical ICA up to a *single* permutation matrix that is common

D. Lahat and Ch. Jutten are with GIPSA-Lab, UMR CNRS 5216, Grenoble Campus, BP46, F-38402 Saint-Martin-d'Hères, France. email:{Dana.Lahat, Christian.Jutten}@gipsa-lab.grenoble-inp.fr.

This work is supported by the project CHESS, 2012-ERC-AdG-320684. GIPSA-Lab is a partner of the LabEx PERSYVAL-Lab (ANR–11-LABX-0025).

to all sets of measurements [4]. Li et al. [5] have shown that the IVA framework, which they termed JBSS, provides sufficient constraints for identifying real Gaussian stationary processes that had been mixed by an invertible matrix, a problem that is ill-posed with classical BSS/ICA, when each mixture is processed separately [31]. This observation, that coupled matrix factorizations enjoy more relaxed uniqueness conditions, finds its tensor counterpart in [23]. Li et al. [32] have shown that JBSS can be reformulated as a coupled matrix diagonalization problem that minimizes a quadratic criterion, and solved by exploiting either second- or higher-order statistics (see also [33]). Recently, JBSS algorithms that minimize the maximum likelihood (ML), mutual information (MI) and entropy have been proposed [34], [35]. When only two data sets are involved, second-order statistics (SOS) JBSS amounts to CCA and can be solved in closed-form using generalized eigenvalue decomposition (GEVD) [36, Chapter 12]. A comprehensive theoretical analysis of IVA can be found in [35] and references therein.

Considering the growing evidence of IVA as a helpful tool in various applications such as multiset data analysis [5], [32], [35], hyperscanning [37] and dynamic systems [38], and the fact that natural signals are often better modeled as multidimensional, it is only natural to take advantage of the benefits of both.

The JISA model, which is the core of this paper, and a SOS-based relative gradient (RG) algorithm that achieves the optimal separation in the presence of real Gaussian data, were first presented in [6]. A Newton-based algorithm that is based on the error analysis in this paper has recently been presented in [39]. We mention that a gradient algorithm that performs JISA based on the multivariate Laplace distribution has recently been proposed in [40]. The novelty and contribution of this paper is in providing a comprehensive theoretical analysis to the SOS approach, including closed-form expressions for the mean square error (MSE), Fisher information matrix (FIM) and Cramér-Rao lower bound (CRLB), as well as proposing a new algebraic formalization that leads to a new, though suboptimal, JISA algorithm.

In this paper, we adopt the approach that was used in [19], [41] to analyse the performance of non-stationary multidimensional BSS. Although these two models are essentially different, exploiting disjoint types of diversity: non-stationarity vs. multiset [35], [42] [43, Section III], using the same approach allows some interesting similarities and analogies between the two models to be manifested. In order to complete the picture, we discuss in this paper a model that can exploit these two types of diversity simultaneously.

The following notations and conventions are used throughout this paper. Regular lowercase, bold lowercase and bold uppercase letters denote scalars, vectors and matrices, respectively. Regular uppercase letters denote functions or operators; calligraphic uppercase letters denote sets. For simplicity, we assume that all values are real. Trace is denoted by $\mathrm{tr}\{\cdot\}$; $(\cdot)^\top$ denotes transpose. $\mathbf{A}^{-\top} = (\mathbf{A}^{-1})^\top$ whenever the inverse exists. $\mathrm{vec}\{\cdot\}$ denotes the operator that stacks the columns of a $P \times Q$ matrix into a $PQ \times 1$ vector. The direct sum of K rectangular matrices $\mathbf{M}^{[k]}$ is denoted by $\oplus_{k=1}^{K}\mathbf{M}^{[k]}$ and yields a block-diagonal matrix $\begin{bmatrix} \mathbf{M}^{[1]} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}^{[K]} \end{bmatrix}$ with $\mathbf{M}^{[k]}$ as its kth diagonal element. The operator $\mathrm{bdiag}_{\mathbf{b}}\{\mathbf{M}\}$, given a $P \times P$ matrix $\mathbf{M}$ and a vector $\mathbf{b}$ of positive integers that sum up to P, extracts from $\mathbf{M}$ a block-diagonal matrix with block-pattern $\mathbf{b}$, and zeroes the off-diagonal blocks. $\mathrm{bdiag}_{\mathbf{b}}^{-1}\{\cdot\}$ stands for $(\mathrm{bdiag}_{\mathbf{b}}\{\cdot\})^{-1}$. $\mathcal{B}_{\mathbf{b}}$ denotes the subspace of all invertible block-diagonal matrices with block-pattern $\mathbf{b}$. $\mathbf{0}$ denotes a one- or two-dimensional array of zeros. $\mathbf{1}_P$ denotes a $P \times 1$ vector of ones. $\mathbf{I}_P$ stands for the $P \times P$ identity matrix, with dimensions that are omitted if they are implicit. $E\{\cdot\}$ denotes expectation. $\mathrm{Cov}(\mathbf{a}) = E\{\mathbf{a}\mathbf{a}^\top\}$, $\mathrm{Cov}(\mathbf{a},\mathbf{b}) = E\{\mathbf{a}\mathbf{b}^\top\}$ for any stochastic vectors $\mathbf{a},\mathbf{b}$ with $E\{\mathbf{a}\} = \mathbf{0}$. $\|\cdot\|$ denotes the Frobenius norm; δ_{ij} denotes the Kronecker delta. The Kronecker product is denoted by $\otimes$. Let $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ denote the (i,j)th $m_i \times n_j$ and $p_i \times q_j$ blocks of partitioned matrices $\mathbf{A}$ and $\mathbf{B}$, respectively. Then, the Khatri-Rao product for partitioned matrices [44], [45] is defined as $\mathbf{A} \boxtimes \mathbf{B} = \begin{bmatrix} \mathbf{A}_{11}\otimes\mathbf{B}_{11} & \mathbf{A}_{12}\otimes\mathbf{B}_{12} & \cdots \\ \mathbf{A}_{21}\otimes\mathbf{B}_{21} & \mathbf{A}_{22}\otimes\mathbf{B}_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$, where $\mathbf{A}_{ij} \otimes \mathbf{B}_{ij}$ and $\mathbf{A}\boxtimes\mathbf{B}$ are $m_i p_i \times n_j q_j$ and $(\sum m_i p_i)\times(\sum n_j q_j)$, respectively. $\mathcal{A}\backslash\mathcal{B}$ denotes the difference between two sets. $O(f)$ stands for deterministic terms that are bounded above, up to a constant factor, by f, or zero-mean stochastic terms whose standard deviation is proportional to f or to higher powers thereof; the specific interpretation is implicit in the context.

The rest of this paper is organized as follows. In Section II, we present and define the model that we denote JISA, and formalize it mathematically. In Section III, we present a SOS approach to JISA. Sections II–III recall results from [6], whereas the novelty is in the sections that follow. Section IV provides a theoretical small-error analysis of the proposed approach. Section V briefly discusses the well-posedness of the model. Section VI reformulates JISA as a model-fit problem with a Frobenius norm. This reformulation leads to a coupled tensor decomposition that can exploit also non-stationarity or correlation among samples. Numerical experiments in Section VII validate our theoretical results, and provide a comparison with two related approaches. We conclude our paper in Section VIII.

II. JISA: Model and Problem Formulation

Consider T observations of K vectors $\mathbf{x}^{[k]}(t)$, modeled as

$$\mathbf{x}^{[k]}(t) = \mathbf{A}^{[k]}\mathbf{s}^{[k]}(t) \quad 1 \le t \le T\,,\, 1 \le k \le K\,, \tag{1}$$

where $\mathbf{A}^{[k]}$ are $M \times M$ invertible matrices that may be different $\forall k$, and $\mathbf{x}^{[k]}(t)$ and $\mathbf{s}^{[k]}(t)$ are $M \times 1$ vectors. For fixed k, each mixture (1) corresponds to classical BSS. In IVA, the elements of the $K \times 1$ vector $\mathbf{s}_i^{\mathrm{IVA}}(t) = [s_i^{[1]}(t),\ldots,s_i^{[K]}(t)]^\top$, $i = 1,\ldots,M$, are statistically *dependent* whereas the pairs $(\mathbf{s}_i^{\mathrm{IVA}}(t),\mathbf{s}_j^{\mathrm{IVA}}(t))$ are statistically *independent* for all $i \neq j \in \{1,\ldots,M\}$. Therefore, IVA aims at extracting M mutually independent vector elements (whence its name) from K sets of measurements by exploiting not only the statistical *independence within each set of measurements* but also the *dependence across sets of measurements*.

Given the partition $\mathbf{s}^{[k]}(t) = [\mathbf{s}_1^{[k]\top}(t),\ldots,\mathbf{s}_N^{[k]\top}(t)]^\top$, where $N \le M$, $\mathbf{s}_i^{[k]}(t)$ are $m_i \times 1$ vectors, $m_i \ge 1$, $\sum_{i=1}^{N} m_i =$

M, and the probability density function (pdf) of each m_i-dimensional random vector $\mathbf{s}_i^{[k]}$ irreducible in the sense that it cannot be factorized into a product of non-trivial pdfs, then each mixture (1) represents a single ISA problem. The *model* that we define[1] as JISA corresponds to linking several such ISA problems via the assumption that the elements of the $n_i \times 1$ vector $\mathbf{s}_i(t) = [\mathbf{s}_i^{[1]\top}(t), \ldots, \mathbf{s}_i^{[K]\top}(t)]^\top$, where $n_i = Km_i$, are statistically *dependent* whereas the pairs $(\mathbf{s}_i(t), \mathbf{s}_j(t))$ are statistically *independent* for all $i \neq j \in \{1, \ldots, N\}$. Figure 1 illustrates this model.

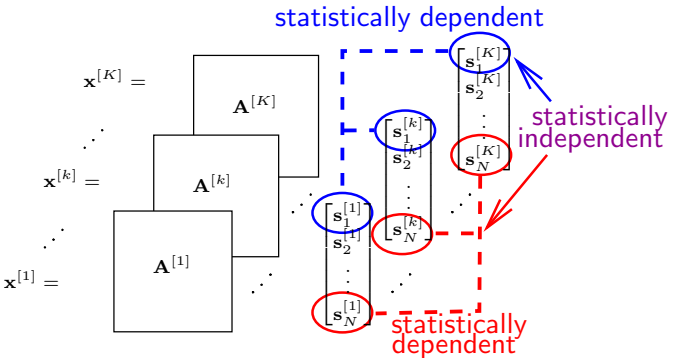


Fig. 1. Diagram of the basic JISA model. Terms with the same subscript are statistically dependent. In all other cases, there is no statistical dependence.

Given $\mathbf{m} = [m_1, \ldots, m_N]^\top$ and a set of observations $\mathcal{X} = \{\mathbf{x}^{[k]}(t)\}_{k=1,t=1}^{K,T}$, the *problem* of JISA is that of finding linear transformations $\mathbf{A}^{-[k]}$ on $\mathcal{X}$, where $\mathbf{A}^{-[k]}$ denotes $(\mathbf{A}^{[k]})^{-1}$, such that the source vectors $\mathbf{s}_1(t), \ldots, \mathbf{s}_N(t)$ are as independent as possible. This notion is given a definite meaning in Section III, where we set up a simple statistical model that, via its likelihood function, yields a quantitative measure of independence.

The above partition of $\mathbf{s}^{[k]}(t)$ induces a corresponding partition in the mixing matrices: $\mathbf{A}^{[k]} = [\mathbf{A}_1^{[k]} | \cdots | \mathbf{A}_N^{[k]}]$ with $\mathbf{A}_i^{[k]}$ the ith $M \times m_i$ column-block of $\mathbf{A}^{[k]}$. The multiplicative model (1) may now be rewritten as a sum of $N \leq M$ *multidimensional components*:

$$\mathbf{x}^{[k]}(t) = \sum_{i=1}^{N} \mathbf{x}_i^{[k]}(t) \tag{2}$$

where the ith $M \times 1$ component vector $\mathbf{x}_i^{[k]}(t)$ is defined as

$$\mathbf{x}_i^{[k]}(t) = \mathbf{A}_i^{[k]} \mathbf{s}_i^{[k]}(t)\,. \tag{3}$$

In a blind context, the component vector $\mathbf{x}_i^{[k]}(t)$ is better defined than the source vector $\mathbf{s}_i^{[k]}(t)$. Indeed, for any invertible $m_i \times m_i$ matrix $\mathbf{Z}_{ii}^{[k]}$, it is impossible to discriminate between the representation of a component $\mathbf{x}_i^{[k]}(t)$ by the pair $(\mathbf{A}_i^{[k]}, \mathbf{s}_i^{[k]}(t))$ and $(\mathbf{A}_i^{[k]} \mathbf{Z}_{ii}^{-[k]}, \mathbf{Z}_{ii}^{[k]} \mathbf{s}_i^{[k]}(t))$. This means that only the column space of $\mathbf{A}_i^{[k]}$, $\mathrm{span}(\mathbf{A}_i^{[k]})$, can be blindly identified. Therefore, JISA is in fact a (joint) subspace estimation problem.

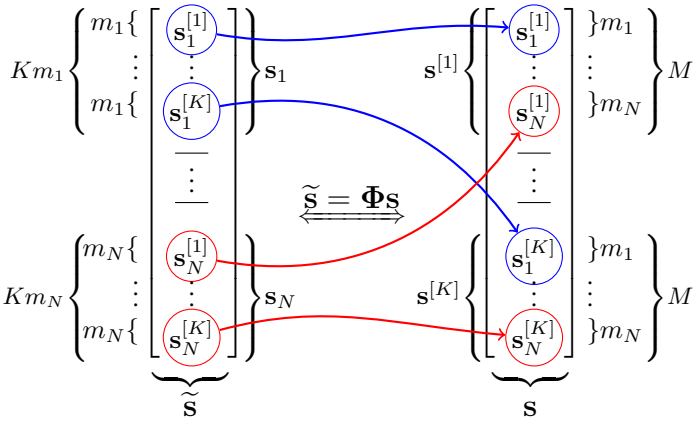
[1]This formulation is sufficiently simple to keep notations and derivations clear and tractable yet at the same time sufficiently rich to encompass the essential properties of JISA that we present.

Further insights can be obtained by stacking all data sets in one vector $\begin{bmatrix} \mathbf{x}^{[1]} \\ \vdots \\ \mathbf{x}^{[K]} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{[1]} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}^{[K]} \end{bmatrix} \begin{bmatrix} \mathbf{s}^{[1]} \\ \vdots \\ \mathbf{s}^{[K]} \end{bmatrix}$, such that (1) rewrites as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{4}$$

where $\mathbf{s}(t) = [\mathbf{s}^{[1]\top}(t), \ldots, \mathbf{s}^{[K]\top}(t)]^\top$ and $\mathbf{x}(t) = [\mathbf{x}^{[1]\top}(t), \ldots, \mathbf{x}^{[K]\top}(t)]^\top$ are $L \times 1$ vectors, $L = KM$, $\mathbf{A} = \oplus_{k=1}^{K} \mathbf{A}^{[k]} \in \mathcal{B}_{\mathbf{k}}$, and $\mathbf{k} = M\mathbf{1}_K$ is the block-pattern of $\mathbf{A}$. Combining (2) and (3) in (4), one obtains that $\mathbf{x}(t) = \sum_{i=1}^{N} \begin{bmatrix} \mathbf{A}_i^{[1]} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_i^{[K]} \end{bmatrix} \begin{bmatrix} \mathbf{s}_i^{[1]}(t) \\ \vdots \\ \mathbf{s}_i^{[K]}(t) \end{bmatrix} = \sum_{i=1}^{N} (\mathbf{I}_K \boxtimes \mathbf{A}_i)\mathbf{s}_i(t) = \sum_{i=1}^{N} \mathbf{x}_i(t)$, where $\mathbf{A}_i \triangleq [\mathbf{A}_i^{[1]} | \cdots | \mathbf{A}_i^{[K]}]$, $\mathbf{x}_i(t) = [\mathbf{x}_i^{[1]\top}(t), \ldots, \mathbf{x}_i^{[K]\top}(t)]^\top$, and $\mathbf{I}_K$ is partitioned columnwise. The latter implies that the N latent components have the same ordering in all the K mixtures. This proves that the ability of IVA to provide a single $M \times M$ permutation matrix (i.e, arbitrary ordering) of the N latent sources in all the involved mixtures (Section I) indeed extends to the multidimensional case. If all the linear transformations on $\mathcal{X}$ that maximize statistical independence between $\mathbf{s}_1(t), \ldots, \mathbf{s}_N(t)$ yield the same unordered set $\{\mathbf{x}_1(t), \ldots, \mathbf{x}_N(t)\}$, we say that the JISA model is unique.

Another useful notation is concatenating the dependent sources $\mathbf{s}_i$ in one vector $\widetilde{\mathbf{s}}(t) = \mathbf{\Phi}\mathbf{s}(t)$, where $\widetilde{\mathbf{s}}(t) = [\mathbf{s}_1^\top(t), \ldots, \mathbf{s}_N^\top(t)]^\top$, and $\mathbf{\Phi}$ is the corresponding $L \times L$ permutation matrix between these two alternative representations:



As we shall see later, it is useful to introduce the separating projectors: these are the $M \times M$ oblique projection matrices $\mathbf{P}_i^{[k]}$ onto $\mathrm{span}(\mathbf{A}_i^{[k]})$ along $\mathrm{span}(\mathbf{A}_j^{[k]})$ $\forall j \neq i$. By definition, they satisfy $\mathbf{P}_i^{[k]} \mathbf{A}_j^{[k]} = \delta_{ij} \mathbf{A}_i^{[k]}$, unaffected if $\mathbf{A}_i^{[k]}$ is changed into $\mathbf{A}_i^{[k]} \mathbf{Z}_{ii}^{-[k]}$ and, most importantly, allow one to write

$$\mathbf{x}_i^{[k]}(t) = \mathbf{P}_i^{[k]} \mathbf{x}^{[k]}(t)\,. \tag{5}$$

Finally, note that if $\mathbf{B}^{[k]} = \mathbf{A}^{-[k]}$ is partitioned into N horizontal $m_i \times M$ blocks $\mathbf{B}_i^{[k]}$, then the rank-m_i ith oblique projection is given by

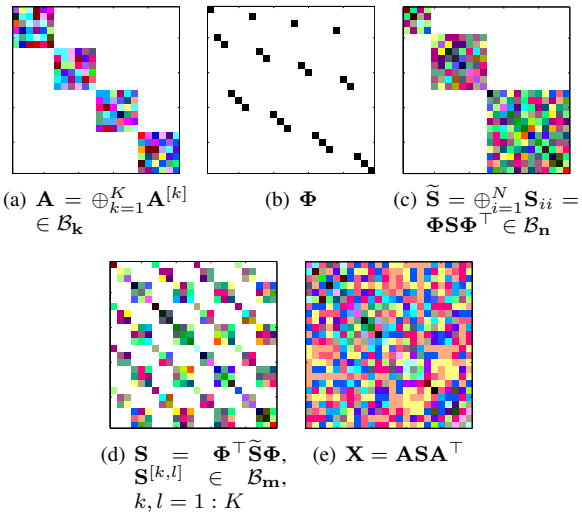
$$\mathbf{P}_i^{[k]} = \mathbf{A}_i^{[k]} \mathbf{B}_i^{[k]}\,. \tag{6}$$

Alternatively, one can define oblique projections such that $\mathbf{P}_i \mathbf{x}(t) = \mathbf{x}_i(t)$. It is easy to verify that $\mathbf{P}_i = \oplus_{k=1}^{K} \mathbf{P}_i^{[k]}$.

In the rest of this paper, we focus on JISA using SOS. Assuming temporally independent and identically distributed (i.i.d.) samples, the model assumptions imply that $\widetilde{\mathbf{S}} \triangleq E\{\widetilde{\mathbf{s}}(t)\widetilde{\mathbf{s}}^\top(t)\} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}_{NN} \end{bmatrix} = \oplus_{i=1}^N \mathbf{S}_{ii} \in \mathcal{B}_{\mathbf{n}}$, where $\widetilde{\mathbf{S}}$ is an $L \times L$ block-diagonal matrix with block-pattern $\mathbf{n} = [n_1, \ldots, n_N]^\top$ and

$$\widetilde{\mathbf{S}} = \mathbf{\Phi}\mathbf{S}\mathbf{\Phi}^\top \in \mathcal{B}_{\mathbf{n}}\,. \tag{7}$$

The (i,j)th $n_i \times n_j$ block of $\widetilde{\mathbf{S}}$ is $\mathbf{S}_{ij} = E\{\mathbf{s}_i(t)\mathbf{s}_j^\top(t)\}$ for $1 \le i,j \le N$. Its empirical counterpart is $\overline{\mathbf{S}}_{ij} = \frac{1}{T}\sum_{t=1}^T \mathbf{s}_i(t)\mathbf{s}_j^\top(t)$. The same statistical assumptions imply that the (k,l)th $M \times M$ block of $\mathbf{S}$ is $\mathbf{S}^{[k,l]} \triangleq E\{\mathbf{s}^{[k]}(t)\mathbf{s}^{[l]\top}(t)\} = \begin{bmatrix} \mathbf{S}_{11}^{[k,l]} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}_{NN}^{[k,l]} \end{bmatrix} = \oplus_{i=1}^N \mathbf{S}_{ii}^{[k,l]} \in \mathcal{B}_{\mathbf{m}}$ for $1 \le k,l \le K$, where $\mathbf{S}_{ij}^{[k,l]} = E\{\mathbf{s}_i^{[k]}(t)\mathbf{s}_j^{[l]\top}(t)\}$ is $m_i \times m_j$ and its empirical counterpart is $\overline{\mathbf{S}}_{ij}^{[k,l]} = \frac{1}{T}\sum_{t=1}^T \mathbf{s}_i^{[k]}(t)\mathbf{s}_j^{[l]\top}(t)$. $\mathbf{S}_{ij}^{[k,l]}$ is thus the (i,j)th block of $\mathbf{S}^{[k,l]}$ as well as the (k,l)th block of $\mathbf{S}_{ij}$. The linear model (4) implies that $\mathbf{X} = \mathbf{A}\mathbf{S}\mathbf{A}^\top$ where $\mathbf{S} = E\{\mathbf{s}(t)\mathbf{s}^\top(t)\}$ and $\mathbf{X} = E\{\mathbf{x}(t)\mathbf{x}^\top(t)\}$. For simplicity, we assume that all $\mathbf{S}_{ii}$ are invertible and do not contain zeros; in practice, this assumption could be relaxed [24], see Section V-B for further details. Typical structures of $\mathbf{A}$, $\mathbf{\Phi}$, $\mathbf{S}$, $\widetilde{\mathbf{S}}$ and $\mathbf{X}$ are illustrated in Figure 2.



(a) $\mathbf{A} = \oplus_{k=1}^K \mathbf{A}^{[k]} \in \mathcal{B}_{\mathbf{k}}$ (b) $\mathbf{\Phi}$ (c) $\widetilde{\mathbf{S}} = \oplus_{i=1}^N \mathbf{S}_{ii} = \mathbf{\Phi}\mathbf{S}\mathbf{\Phi}^\top \in \mathcal{B}_{\mathbf{n}}$

(d) $\mathbf{S} = \mathbf{S}^{[k,l]} \in \mathcal{B}_{\mathbf{m}}$, $k,l = 1:K$ (e) $\mathbf{X} = \mathbf{A}\mathbf{S}\mathbf{A}^\top$

Fig. 2. Typical model parameters used in JISA. In this example, there are $K = 4$ mixtures. In each mixture, there are $N = 3$ components with dimensions (block-pattern) $\mathbf{m} = [1,2,3]^\top$. Hence, $M = \sum_{i=1}^N m_i = 6$, $L = MN = 18$, $\mathbf{n} = K\mathbf{m} = [4,8,12]^\top$, $\mathbf{k} = M\mathbf{1}_K = [6,6,6,6]^\top$. The color scale is arbitrary and different in each subfigure, except (c) and (d).

III. Optimal Component Separation Using Second-Order Statistics

In the following, we consider a Gaussian model in which $\mathbf{s}_i(t) \sim \mathcal{N}(\mathbf{0}_{n_i \times 1}, \mathbf{S}_{ii})$ are mutually independent samples $\forall t \neq t'$. The log-likelihood for the model just described is

$$\log p(\mathcal{X}; \mathcal{A}, \mathbf{S}) \triangleq -T\phi(\mathcal{A}, \mathbf{S}) = \sum_{t=1}^T \log p(\mathbf{x}(t)) \tag{8a}$$

$$= -\frac{1}{2}\sum_{t=1}^T \left(\log\det 2\pi\mathbf{X} + \mathbf{x}^\top(t)\mathbf{X}^{-1}\mathbf{x}(t)\right) \tag{8b}$$

$$= -\frac{T}{2}\left(\log\det 2\pi\mathbf{X} + \mathrm{tr}\{\overline{\mathbf{X}}\mathbf{X}^{-1}\}\right) \tag{8c}$$

$$= -TD(\overline{\mathbf{X}}, \mathbf{X}) - \kappa = -TD(\overline{\mathbf{X}}, \mathbf{A}\mathbf{S}\mathbf{A}^\top) - \kappa \tag{8d}$$

$$= -TD(\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\mathbf{\Phi}^\top, \widetilde{\mathbf{S}}) - \kappa \tag{8e}$$

where $\mathcal{A} = \{\mathbf{A}^{[k]}\}_{k=1}^K$, and $\overline{\mathbf{X}} = \frac{1}{T}\sum_{t=1}^T \mathbf{x}(t)\mathbf{x}^\top(t)$ is the empirical counterpart of $\mathbf{X}$. The second equality in (8a) is due to the assumption of pairwise sample independence for $t \neq t'$. Equation (8b) is due to the Gaussian assumption and (4), which imply $\mathbf{x}(t) \sim \mathcal{N}(\mathbf{0}_{L\times 1}, \mathbf{X})$. Equation (8c) follows from $\mathbf{a}^\top\mathbf{R}\mathbf{a} = \mathrm{tr}\{\mathbf{R}\mathbf{a}\mathbf{a}^\top\}$ for any vector $\mathbf{a}$ and matrix $\mathbf{R}$ of appropriate dimensions. The scalar

$$D(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{2}(\mathrm{tr}\{\mathbf{R}_1\mathbf{R}_2^{-1}\} - \log\det(\mathbf{R}_1\mathbf{R}_2^{-1}) - Q)\,, \tag{9}$$

defined for any two $Q \times Q$ symmetric positive-definite matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, is the Kullback-Leibler divergence (KLD) between the distributions $\mathcal{N}(\mathbf{0}, \mathbf{R}_1)$ and $\mathcal{N}(\mathbf{0}, \mathbf{R}_2)$ [46]. The term $\kappa = \frac{T}{2}(\log\det(2\pi\overline{\mathbf{X}}) + L)$ is irrelevant to the maximization of the likelihood since it depends only on the data and not on the parameters. Equation (8e) follows from (7), (9) and (41d). The derivation of (8) follows similar lines as those used in [41, Section III] and [47, Section 3].

A. Contrast Function

Given the block-diagonal structure of $\widetilde{\mathbf{S}}$, the last step in (8) gives rise to its ML estimate [41, Appendix B]

$$\widehat{\widetilde{\mathbf{S}}}^{\mathrm{ML}} = \mathrm{bdiag}_{\mathbf{n}}\{\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\mathbf{\Phi}^\top\} \tag{10a}$$

$$\widehat{\mathbf{S}}^{\mathrm{ML}} = \mathbf{\Phi}^\top\,\mathrm{bdiag}_{\mathbf{n}}\{\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\mathbf{\Phi}^\top\}\mathbf{\Phi} \tag{10b}$$

where (10b) is due to (7). Note that the result in [47, Section 3.3] is a special case of (10b) when $m_i = 1\ \forall i$. We can now write

$$\max_{\mathbf{S}} \log p(\mathcal{X}; \mathcal{A}, \mathbf{S}) = -TC(\mathbf{A}) + \kappa\,, \tag{11}$$

where in the latter we have defined the *contrast function* [48]

$$C(\mathbf{A}) = D(\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\mathbf{\Phi}^\top, \mathrm{bdiag}_{\mathbf{n}}\{\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\mathbf{\Phi}^\top\})\,. \tag{12}$$

It holds that $D(\mathbf{R}, \mathrm{bdiag}_{\mathbf{b}}\{\mathbf{R}\}) \ge 0$ with equality if and only if (iff) $\mathbf{R} \in \mathcal{B}_{\mathbf{b}}$. Hence, for any positive-definite matrix $\mathbf{R}$, $D(\mathbf{R}, \mathrm{bdiag}_{\mathbf{b}}\{\mathbf{R}\})$ is a measure of the block-diagonality of $\mathbf{R}$. Therefore, minimizing the contrast function[2] (12) amounts to *(approximate) block diagonalization* of $\overline{\mathbf{X}}$ by a permuted block-diagonal matrix $\mathbf{\Phi}\mathbf{A}^{-1}$.

[2]We assume that an optimum exists.

B. Estimating Equations

The next step is obtaining an ML estimate of $\mathcal{A}$. For this purpose, we calculate the derivative of the minus log-likelihood $\phi(\mathcal{A},\mathbf{S})$ with respect to (w.r.t.) each $\mathbf{A}^{[k]}$ separately, for fixed $\mathbf{S}$ and $\mathcal{A}\backslash\mathbf{A}^{[k]}$, as we now explain. Consider a relative variation $\mathbf{A}^{[k]} \to \mathbf{A}^{[k]}(\mathbf{I}+\boldsymbol{\delta}^{[k]})^{-1}$, where $\boldsymbol{\delta}^{[k]}$ is $M\times M$ and has arbitrary values but such that $\mathbf{I}+\boldsymbol{\delta}^{[k]}$ is invertible[3]. Then, the first-order variation of $\phi(\mathcal{A},\mathbf{S})$, when $\mathbf{A}^{[k]}$ is replaced by $\mathbf{A}^{[k]}(\mathbf{I}+\boldsymbol{\delta}^{[k]})^{-1}$ and the other mixing matrices remain unchanged, can always be expressed by the Taylor expansion

$$\begin{aligned}\phi(\{\mathcal{A}\setminus\mathbf{A}^{[k]},\mathbf{A}^{[k]}(\mathbf{I}+\boldsymbol{\delta}^{[k]})^{-1}\},\mathbf{S}) &= \phi(\mathcal{A},\mathbf{S}) \\ &+ \mathrm{tr}\{(\nabla\phi^{[k]}(\mathcal{A},\mathbf{S}))^\top\boldsymbol{\delta}^{[k]}\} + \text{higher-order terms in } \boldsymbol{\delta}^{[k]}\,,\end{aligned} \tag{13}$$

where $\nabla\phi^{[k]}(\mathcal{A},\mathbf{S})$ denotes the $M\times M$ RG of $\phi(\mathcal{A},\mathbf{S})$ w.r.t. $\mathbf{A}^{[k]}$. Equation (13) follows from the definition of the RG in [49, Section III.C]. Derivation similar to [41, Section III.D] yields

$$\nabla\phi^{[k]}(\mathcal{A},\mathbf{S}) = \mathbf{J}_k^\top\mathbf{S}^{-1}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\mathbf{J}_k - \mathbf{I}_M\,, \tag{14}$$

where $\mathbf{J}_k$ is the kth $L\times M$ block of $\mathbf{I}_L=[\mathbf{J}_1,\ldots,\mathbf{J}_K]$. The K terms $\nabla\phi^{[k]}(\mathcal{A},\mathbf{S})$ in (14) can be collected into

$$\begin{aligned}\nabla\phi(\mathcal{A},\mathbf{S}) &\triangleq \sum_{k=1}^K \mathbf{J}_k\nabla\phi^{[k]}(\mathcal{A},\mathbf{S})\mathbf{J}_k^\top = \oplus_{k=1}^K\nabla\phi^{[k]}(\mathcal{A},\mathbf{S}) \\ &= \mathrm{bdiag}_{\mathbf{k}}\{\mathbf{S}^{-1}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\} - \mathbf{I}_L\,.\end{aligned} \tag{15}$$

It can be shown that the first-order variation of $C(\mathbf{A})$ w.r.t. $\mathbf{A}^{[k]}$, derived similarly to (13)–(14), obeys

$$\nabla C^{[k]}(\mathbf{A}) = \nabla\phi^{[k]}(\mathcal{A},\mathbf{S})\big|_{\mathbf{S}=\widehat{\mathbf{S}}^{\mathrm{ML}}}\,. \tag{16}$$

Given (15), (16), (10b) and $\nabla C(\mathbf{A}) = \oplus_{k=1}^K\nabla C^{[k]}(\mathbf{A})$, we can now write

$$\begin{aligned}&\nabla C(\mathbf{A}) = \nabla\phi(\mathcal{A},\mathbf{S})|_{\mathbf{S}=\widehat{\mathbf{S}}^{\mathrm{ML}}} = \\ &\mathrm{bdiag}_{\mathbf{k}}\{\boldsymbol{\Phi}^\top\,\mathrm{bdiag}_{\mathbf{n}}^{-1}\{\boldsymbol{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\boldsymbol{\Phi}^\top\}\boldsymbol{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\} - \mathbf{I}\,.\end{aligned} \tag{17}$$

Values of $\mathcal{A}$ that maximize the likelihood and thus minimize $C(\mathbf{A})$ also satisfy $\nabla C(\mathbf{A})=\mathbf{0}$. Henceforth, matrices that satisfy the *estimating equations*

$$\mathrm{bdiag}_{\mathbf{k}}\{\boldsymbol{\Phi}^\top\,\mathrm{bdiag}_{\mathbf{n}}^{-1}\{\boldsymbol{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\boldsymbol{\Phi}^\top\}\boldsymbol{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\} = \mathbf{I} \tag{18}$$

are denoted $\widehat{\mathbf{A}}^{[k]\mathrm{ML}}$. The associated oblique projections (6) are denoted $\widehat{\mathbf{P}}_i^{[k]\mathrm{ML}}$. The corresponding component estimates are given by

$$\widehat{\mathbf{x}}_i^{[k]\mathrm{ML}}(t) \triangleq \widehat{\mathbf{P}}_i^{[k]\mathrm{ML}}\mathbf{x}^{[k]}(t)\,, \tag{19}$$

which follows from (5).

[3]Matrix $\mathbf{A}$ is block-diagonal *by definition* and thus there is no meaning to perturbing its off-block-diagonal entries. This is the bifurcation point from which the derivation takes a different path than that in [41].

C. Figure of Merit: Mean Square Error

Our goal is component separation. Therefore, the problem of JISA consists in estimating $\mathbf{x}_i^{[k]}(t)$ given only $\mathcal{X}$ and $\mathbf{m}$. We define the MSE as the figure of merit in the estimation of $\mathbf{x}_i^{[k]}(t)$,

$$\widehat{\mathrm{MSE}}_i^{[k]} = \frac{1}{T}\sum_{t=1}^T\|\widehat{\mathbf{x}}_i^{[k]}(t)-\mathbf{x}_i^{[k]}(t)\|^2\,. \tag{20}$$

Alternatively, we may be interested in the normalized MSE in the estimation of $\mathbf{x}_i(t)$,

$$\widehat{\mathrm{MSE}}_i = \frac{1}{\sigma_i^2}\frac{1}{T}\sum_{t=1}^T\|\widehat{\mathbf{x}}_i(t)-\mathbf{x}_i(t)\|^2 = \frac{1}{\sigma_i^2}\sum_{k=1}^K\widehat{\mathrm{MSE}}_i^{[k]}\,, \tag{21}$$

where $\sigma_i^2 = E\{\|\mathbf{x}_i(t)\|^2\}$. For Gaussian data, estimates of $\mathbf{x}_i(t)$ obtained via (19) from matrices that satisfy (18) achieve asymptotically (i.e., $T\to\infty$) the minimal mean square error (MMSE).

IV. Error Analysis

We now turn to the error analysis of our model. This will lead us to a closed-form expression for the FIM and CRLB in the estimation of the oblique projections and to the MSE in component estimation.

A. Error Decomposition

A difficulty in the error analysis of blind subspace estimation stems from the inability to characterize the error in the mixing matrices, due to severe indeterminacies they suffer from (Section II). We thus begin by defining convenient error terms. In order to focus on well-defined quantities, we consider the errors

$$\delta\mathbf{P}_i^{[k]} \triangleq \widehat{\mathbf{P}}_i^{[k]} - \mathbf{P}_i^{[k]} \tag{22}$$

in $\widehat{\mathbf{P}}_i^{[k]}$, the estimates of the oblique projectors $\mathbf{P}_i^{[k]}$. Accordingly, the estimate of $\mathbf{x}_i^{[k]}(t)$ is $\widehat{\mathbf{x}}_i^{[k]}(t) = \widehat{\mathbf{P}}_i^{[k]}\mathbf{x}^{[k]}(t) = \mathbf{x}_i^{[k]}(t) + \delta\mathbf{P}_i^{[k]}\mathbf{x}^{[k]}(t)$, which follows from (5), (19) and (22). Consequently, the component estimation error is given by

$$\widehat{\mathbf{x}}_i^{[k]}(t) - \mathbf{x}_i^{[k]}(t) = \delta\mathbf{P}_i^{[k]}\mathbf{x}^{[k]}(t)\,. \tag{23}$$

Equation (20) can now be rewritten as

$$\begin{aligned}\widehat{\mathrm{MSE}}_i^{[k]} &= \frac{1}{T}\sum_{t=1}^T\|\delta\mathbf{P}_i^{[k]}\mathbf{x}^{[k]}(t)\|^2 \\ &= \mathrm{tr}\{(\overline{\mathbf{X}}^{[k,k]}\otimes\mathbf{I}_M)\mathrm{vec}\{\delta\mathbf{P}_i^{[k]}\}\mathrm{vec}^\top\{\delta\mathbf{P}_i^{[k]}\}\}\end{aligned} \tag{24}$$

where the last equality uses $\|\mathbf{a}\|^2=\mathrm{tr}\{\mathbf{a}\mathbf{a}^\top\}$ and Property A.1 in Appendix A. Matrices $\overline{\mathbf{X}}^{[k,l]}$ and $\mathbf{X}^{[k,l]}$ denote the (k,l)th blocks, according to block-partition $\mathbf{k}$, of $\overline{\mathbf{X}}$ and $\mathbf{X}$, respectively. It can be shown (Appendix B) that asymptotically,

$$\begin{aligned}\mathrm{MSE}_i^{[k]} &\triangleq E\{\widehat{\mathrm{MSE}}_i^{[k]}\} \\ &= \mathrm{tr}\{(\mathbf{X}^{[k,k]}\otimes\mathbf{I}_M)\,\mathrm{Cov}(\mathrm{vec}\{\delta\mathbf{P}_i^{[k]}\})\} + O(\tfrac{1}{T^{3/2}})\,.\end{aligned} \tag{25}$$

In the following, we set out to obtain a closed-form expression for $\mathrm{Cov}(\mathrm{vec}\{\delta\mathbf{P}_i^{[k]}\})$ as a function only of the model parameters, that will conclude the derivation of the MSE.

B. First-Order Approximation of $\delta\mathbf{P}_i^{[k]}$

In general, any estimate or approximation of $\mathbf{A}^{[k]}$ can be rewritten as a product of $\mathbf{A}^{[k]}$ and some perturbation matrix. In addition, as explained in Appendix D, the contrast function (12) is invariant to right-multiplying each $\mathbf{A}^{[k]}$ by any $\mathbf{\Lambda}^{[k]} \in \mathcal{B}_{\mathbf{m}}$. Hence, the most general form of the minimizer of (12) can be formulated as

$$\widehat{\mathbf{A}}^{[k]} = \mathbf{A}^{[k]}(\mathbf{I}_M + \boldsymbol{\mathcal{E}}^{[k]})^{-1}\mathbf{\Lambda}^{[k]}\,, \tag{26}$$

where the $M \times M$ matrix $\boldsymbol{\mathcal{E}}^{[k]}$ reflects the *relative* change in $\mathbf{A}^{[k]}$, up to the scale ambiguity which is represented by $\mathbf{\Lambda}^{[k]}$. In Appendix E we show that

$$\boldsymbol{\delta}\mathbf{P}_i^{[k]} = \mathbf{A}^{[k]}\Big(\sum_{j\neq i}\mathbf{E}_i\boldsymbol{\mathcal{E}}_{ij}^{[k]}\mathbf{E}_j^\top - \mathbf{E}_j\boldsymbol{\mathcal{E}}_{ji}^{[k]}\mathbf{E}_i^\top\Big)\mathbf{A}^{-[k]} + \text{higher-order terms in } \boldsymbol{\mathcal{E}}^{[k]} \tag{27}$$

where $\mathbf{E}_i$ is the ith $M \times m_i$ block of $\mathbf{I}_M = [\mathbf{E}_1, \ldots, \mathbf{E}_N]$. The $m_i \times m_j$ matrix $\boldsymbol{\mathcal{E}}_{ij}^{[k]}$ denotes the (i,j)th block of $\boldsymbol{\mathcal{E}}^{[k]}$, according to partition $\mathbf{m}$. Since $\mathbf{\Lambda}^{[k]}$ has vanished from (27), we can proceed with our error analysis without worrying about the scale ambiguity.

C. Influence Function

In order to evaluate the covariance of the error terms, we begin by establishing the first-order expansion of $\boldsymbol{\mathcal{E}}_{ij}^{[k]}$ in terms of the sample covariance matrices. In this section, we develop the error analysis in the regime of small errors; that is, we analyze the error terms $\mathcal{E} = \{\boldsymbol{\mathcal{E}}^{[k]}\}_{k=1}^K$ at first-order in $\overline{\mathbf{S}}_{ij}^{[k,l]}$, when asymptotic conditions (Section III-C) hold. Our source separation method is based on the key assumption that $\mathbf{S}_{ij}^{[k,l]} = E\{\overline{\mathbf{S}}_{ij}^{[k,l]}\} = \mathbf{0}_{m_i\times m_j}$ for $j \neq i$ and any k, l. However, because of finite sample size, its empirical counterpart, $\overline{\mathbf{S}}_{ij}^{[k,l]} \neq \mathbf{0}_{m_i\times m_j}$, does not hold.

The first-order expansion of the estimating equations (18) yields (see Appendix F) $KN(N-1)$ equations that can be written pairwise, for each $i \neq j$ and all k, as

$$\begin{aligned} -[\mathbf{S}_{ii}^{-1}\overline{\mathbf{S}}_{ij}]_{kk} &= \sum_{l=1}^K [\mathbf{S}_{ii}^{-1}]_{kl}\boldsymbol{\mathcal{E}}_{ij}^{[l]}\mathbf{S}_{jj}^{[l,k]} + (\boldsymbol{\mathcal{E}}_{ji}^{[k]})^\top + O(\tfrac{1}{T}) \\ -[\mathbf{S}_{jj}^{-1}\overline{\mathbf{S}}_{ji}]_{kk} &= \sum_{l=1}^K [\mathbf{S}_{jj}^{-1}]_{kl}\boldsymbol{\mathcal{E}}_{ji}^{[l]}\mathbf{S}_{ii}^{[l,k]} + (\boldsymbol{\mathcal{E}}_{ij}^{[k]})^\top + O(\tfrac{1}{T}) \end{aligned} \tag{28}$$

where $[\cdot]_{kl}$ stands for the (k,l)th block of the term in brackets in the appropriate partition. The pairwise symmetry of the equations in (28) highlights the fact that asymptotically, for each pair of components $(i \neq j)$, the error terms $\{\boldsymbol{\mathcal{E}}_{ij}^{[k]}, \boldsymbol{\mathcal{E}}_{ji}^{[k]}\}_{k=1}^K$ are related to the corresponding pair of matrices $(\overline{\mathbf{S}}_{ij}, \overline{\mathbf{S}}_{ji})$ that represents the error in the decorrelation of different groups of dependent sources. This type of pairwise decoupling arises naturally in the asymptotic analysis of source separation models that exploit pairwise independence, for example [50, Theorem 11] [19], [41], [51], [52].

Using the $\text{vec}\{\cdot\}$ operator, (28) can be rewritten, for each pair $i \neq j$, as

$$-\mathbf{g} = \boldsymbol{\mathcal{H}}\mathbf{e} + O(\tfrac{1}{T}) \tag{29}$$

where $\mathbf{e}$ and $\mathbf{g}$ are $2Km_im_j \times 1$ vectors,

$$\mathbf{e} = \begin{bmatrix}\mathbf{e}_{ij}\\ \mathbf{e}_{ji}\end{bmatrix}\,,\ \mathbf{e}_{ij} = \begin{bmatrix}\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}^{[1]}\}\\ \vdots \\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}^{[K]}\}\end{bmatrix}\,, \tag{30}$$

$$\mathbf{g} = \begin{bmatrix}\mathbf{g}_{ij}\\ \mathbf{g}_{ji}\end{bmatrix}\,,\ \mathbf{g}_{ij} = \begin{bmatrix}\text{vec}\{[\mathbf{S}_{ii}^{-1}\overline{\mathbf{S}}_{ij}]_{11}\}\\ \vdots \\ \text{vec}\{[\mathbf{S}_{ii}^{-1}\overline{\mathbf{S}}_{ij}]_{KK}\}\end{bmatrix} \tag{31}$$

and

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix}\mathbf{S}_{jj}\boxtimes\mathbf{S}_{ii}^{-1} & \mathbf{I}_K\otimes\boldsymbol{\mathcal{T}}_{m_j,m_i}\\ \mathbf{I}_K\otimes\boldsymbol{\mathcal{T}}_{m_i,m_j} & \mathbf{S}_{ii}\boxtimes\mathbf{S}_{jj}^{-1}\end{bmatrix} \tag{32}$$

is a $2Km_im_j \times 2Km_im_j$ matrix.

$$\mathbf{S}_{jj}\boxtimes\mathbf{S}_{ii}^{-1} = \begin{bmatrix}\mathbf{S}_{jj}^{[1,1]}\otimes[\mathbf{S}_{ii}^{-1}]_{11} & \cdots & \mathbf{S}_{jj}^{[1,K]}\otimes[\mathbf{S}_{ii}^{-1}]_{11}\\ \vdots & & \vdots\\ \mathbf{S}_{jj}^{[K,1]}\otimes[\mathbf{S}_{ii}^{-1}]_{K1} & \cdots & \mathbf{S}_{jj}^{[K,K]}\otimes[\mathbf{S}_{ii}^{-1}]_{KK}\end{bmatrix}$$

is a $Km_im_j \times Km_im_j$ matrix partitioned into blocks according to $m_im_j\mathbf{1}_K$, whose (k,l)th block is $\mathbf{S}_{jj}^{[k,l]}\otimes[\mathbf{S}_{ii}^{-1}]_{kl}$ and has dimensions $m_im_j \times m_im_j$. In the transition from (28) to (29), (30), (31) and (32) we have used the identities (41) in Appendix A. In (32) we introduce the commutation matrix $\boldsymbol{\mathcal{T}}_{P,Q} \in \mathbb{R}^{PQ\times PQ}$, where $\text{vec}\{\mathbf{M}^\top\} = \boldsymbol{\mathcal{T}}_{P,Q}\text{vec}\{\mathbf{M}\}$ for any $\mathbf{M} \in \mathbb{R}^{P\times Q}$ [53]. More properties of the commutation matrix can be found in Appendix A. Assuming that $\boldsymbol{\mathcal{H}}$ is invertible[4], we rewrite (29) as

$$\mathbf{e} = -\boldsymbol{\mathcal{H}}^{-1}\mathbf{g} + O(\tfrac{1}{T}) \quad i \neq j\,. \tag{33}$$

Equation (33) shows how the empirical correlation between the sources, that is, the fact that $\overline{\mathbf{S}}_{ij}$ is non-zero in finite sample size, results in non-zero terms $\mathcal{E}$. Equation (33) is the desired first-order expression for the error terms in (27).

D. Closed-Form Expressions for $\text{Cov}(\text{vec}\{\boldsymbol{\delta}\mathbf{P}_i^{[k]}\})$ and MSE

The first step in expressing $\text{Cov}(\text{vec}\{\delta\mathbf{P}_i^{[k]}\})$ as a function of the model parameters is vectorizing (27). Using identity (41c) we obtain

$$\text{vec}\{\boldsymbol{\delta}\mathbf{P}_i^{[k]}\} = (\mathbf{A}^{-[k]\top}\otimes\mathbf{A}^{[k]})\sum_{j\neq i}^N\Big((\mathbf{E}_j\otimes\mathbf{E}_i)\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}^{[k]}\} - (\mathbf{E}_i\otimes\mathbf{E}_j)\text{vec}\{\boldsymbol{\mathcal{E}}_{ji}^{[k]}\}\Big) + O(\tfrac{1}{T})\,. \tag{34}$$

The covariance of $\text{vec}\{\boldsymbol{\delta}\mathbf{P}_i^{[k]}\}$ can be expressed as

$$\begin{aligned}\text{Cov}(\text{vec}\{\boldsymbol{\delta}\mathbf{P}_i^{[k]}\}) &= (\mathbf{A}^{-[k]\top}\otimes\mathbf{A}^{[k]})\\ &(\mathbf{M}_{11}+\mathbf{M}_{12}+\mathbf{M}_{21}+\mathbf{M}_{22})(\mathbf{A}^{-[k]}\otimes\mathbf{A}^{[k]\top}) + O(\tfrac{1}{T^{3/2}})\end{aligned} \tag{35}$$

where

$$\mathbf{M} \triangleq \begin{bmatrix}\mathbf{M}_{11}\mathbf{M}_{12}\\ \mathbf{M}_{21}\mathbf{M}_{22}\end{bmatrix} \triangleq \sum_{j\neq i}^{N_c}\mathbf{F}_{ij}\,\text{Cov}\left(\begin{bmatrix}\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}^{[k]}\}\\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}^{[k]}\}\end{bmatrix}\right)\mathbf{F}_{ij}^\top\,, \tag{36}$$

[4]In this paper, we assume that $\boldsymbol{\mathcal{H}}$ in invertible. The invertibility of $\boldsymbol{\mathcal{H}}$ is associated with the uniqueness of the model [24], see Section V-B for further details.

$$\mathbf{F}_{ij} \triangleq \begin{bmatrix} \mathbf{E}_j \otimes \mathbf{E}_i & \mathbf{0} \\ \mathbf{0} & -\mathbf{E}_i \otimes \mathbf{E}_j \end{bmatrix} \text{ and}$$

$$\mathrm{Cov}(\mathbf{e}) = \boldsymbol{\mathcal{H}}^{-1}\,\mathrm{Cov}(\mathbf{g})\boldsymbol{\mathcal{H}}^{-1} = \frac{1}{T}\boldsymbol{\mathcal{H}}^{-1} + O(\tfrac{1}{T^{3/2}})\,, \quad (37)$$

as we now explain. Equation (37) follows from Appendix G, where we show that $\mathrm{Cov}(\mathbf{g}) = \frac{1}{T}\boldsymbol{\mathcal{H}}$. Equation (37) fully characterizes the covariance terms in (36). The asymptotic error term in (37) and (35), as well as the fact that there is summation only over one index (j) in (36), follow from Appendix G, which implies, combined with (33), that the pairs $(\boldsymbol{\mathcal{E}}_{ij}^{[k]}, \boldsymbol{\mathcal{E}}_{ai}^{[k]})$ and $(\boldsymbol{\mathcal{E}}_{ij}^{[k]}, \boldsymbol{\mathcal{E}}_{ia}^{[k]})$ are asymptotically uncorrelated $\forall a \neq j$. We have thus obtained a closed-form expression (35) for the covariance of all the entries of $\delta\mathbf{P}_i^{[k]}$ that is a function only of the model parameters $\mathcal{A}$, $\mathbf{S}$ and $\mathbf{m}$ and that is invariant to the arbitrary scaling between $\mathcal{A}$ and $\mathbf{S}$. This expression can be used in (25) for a closed-form expression of the MSE. Further simplification of the MSE can be obtained by using property (41d) and the block-diagonal structure of $\mathbf{S}^{[k,k]}$. Hence,

$$\begin{aligned}\mathrm{MSE}_i^{[k]} &= \mathrm{tr}\big\{(\mathbf{X}^{[k,k]} \otimes \mathbf{I}_M)(\mathbf{A}^{-[k]\top} \otimes \mathbf{A}^{[k]}) \\ &\quad (\mathbf{M}_{11} + \mathbf{M}_{22})(\mathbf{A}^{-[k]} \otimes \mathbf{A}^{[k]\top})\big\} + O(\tfrac{1}{T^{3/2}}) \quad (38)\end{aligned}$$

where the terms that depend on $\mathbf{M}_{12}$ and $\mathbf{M}_{21}$, defined in (36), vanish.

E. FIM, CRLB and MMSE

For samples that follow the Gaussian model in Section II, the results in this section have the following interpretation. Equation (35) is the asymptotically achievable CRLB on the estimation of $\mathbf{P}_i^{[k]}$, and its inverse is the FIM. From (19), equation (38) is the MMSE in the estimation of $\mathbf{x}_i^{[k]}(t)$.

We point out that all the derivations in Section IV and the related appendices rely only on SOS and thus hold also for non-Gaussian observations. That is, (35) and (38) still reflect the error covariance and MSE if we apply the methods in Section III; however, the CRLB, FIM and MMSE interpretation no longer applies.

V. Well-posedness of the JISA SOS Model

We now discuss conditions under which blind identification of the component subspaces is possible.

A. Degrees of Freedom

Let us compare the number of degrees of freedom in the model with the number of constraints in the data, in a similar manner as in [41, Section V.A]. The data are represented by a symmetric $L \times L$ matrix, such that the model tries to fit $N_{\text{data}} = \frac{1}{2}L(L-1)$ scalar numbers. The model consists of K $M \times M$ mixing matrices and N $n_i \times n_i$ source covariance matrices. These provide $N_{\text{model}} = K(M^2 - \sum_{i=1}^N m_i^2) + \frac{1}{2}\sum_{i=1}^N n_i(n_i-1)$ effective free scalar parameters, when scale ambiguities ($\mathbf{Z}^{[k]} \in \mathcal{B}_{\mathbf{m}}$ in Section II) are taken into account. It is immediate to verify that

$$N_{\text{data}} - N_{\text{model}} = \frac{1}{2}(K-2)(M^2 - \sum_{i=1}^N m_i^2) \quad (39)$$

Hence, as soon as $K \geq 2$, there are as many (or more) distinct data values as free parameters in the model. The same calculation shows that imposing statistical independence between all pairs $(\mathbf{s}_i^{[k]}, \mathbf{s}_i^{[k'\neq k]})$ yields a model that is never blindly identifiable using SOS. This result is not surprising, since such a model amounts to K separate BSS/ICA problems.

B. Uniqueness and Identifiability

The previous argument makes it plausible that for randomly chosen source covariance matrices, the component subspaces can be uniquely identified. In fact, and generalizing IVA [54], the uniqueness of the JISA model can be preserved even if not all entries of $\mathbf{s}_i(t)$ are mutually statistically dependent [24]; further discussion of this point is beyond the scope of this paper. In the following, we assume that the uniqueness conditions are satisfied. It is only for the simplicity of presentation that, in this paper, we assume that all corresponding components are mutually dependent, i.e., $\mathbf{S}_{ii}$ do not contain zeros. Since the mixing matrices are assumed to be invertible, uniqueness of the decomposition implies identifiability of the model.

VI. Alternative Algebraic Representation of JISA

In Section III-A we established that JISA amounts to (approximate) block-diagonalization of the (sample) covariance of the observations by a permuted block-diagonal matrix, where the permutation and the block structure are assumed to be known. Due to the invariance of the KLD to rotation of its parameters, the contrast function can be rewritten as $C(\mathbf{A}) = D(\overline{\mathbf{X}}, \mathbf{A}\boldsymbol{\Phi}^\top\,\mathrm{bdiag}_{\mathbf{n}}\{\boldsymbol{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\top}\boldsymbol{\Phi}\mathbf{A}^\top\}\boldsymbol{\Phi}\mathbf{A}^\top)$. Relaxing the measure of divergence from KLD to the Frobenius norm and using (7), $\mathcal{A}$ can now be approximated from

$$\min_{\mathcal{A},\boldsymbol{\Sigma}} \sum_{k=1}^K \sum_{l=1}^K \|\overline{\mathbf{X}}^{[k,l]} - \mathbf{A}^{[k]}\boldsymbol{\Sigma}^{[k,l]}\mathbf{A}^{[l]\top}\|^2 \quad (40)$$

where $\boldsymbol{\Sigma}^{[k,l]} = \mathrm{bdiag}_{\mathbf{m}}\{\mathbf{A}^{-[k]}\overline{\mathbf{X}}^{[k,l]}\mathbf{A}^{-[l]\top}\}$ is the (k,l)th block of $\boldsymbol{\Sigma}$. Note that $E\{\boldsymbol{\Sigma}^{[k,l]}\} = \mathbf{S}^{[k,l]}$. Therefore, both criteria, (40) and (12), achieve the same optimum (zero) for infinite sample size. Equation (40) is not the only approximation to (12): other suboptimal model-fit criteria are also possible, see, e.g., [55] for a recent review. Consistently with Section II, we require that the set of block-diagonal matrices $\{\mathbf{S}^{[k,l]}\}_{k,l=1}^K$ be irreducible in the sense that it cannot be further diagonalized into smaller blocks by any coupled linear transformation of the form $\{\mathbf{T}^{-[k]}\mathbf{S}^{[k,l]}\mathbf{T}^{-[l]\top}\}_{k,l=1}^K$. Equation (40) generalizes [32], where $m_i = 1\ \forall i$, to blocks of arbitrary size. Equation (40) can be interpreted as a (approximate) coupled block diagonalization that minimizes the squared Frobenius norm. As such, (40) falls within the framework of structured data fusion (SDF) [56] and can be solved, in a straightforward model-fit approach, using Tensorlab [57].

We now briefly discuss a generalization that highlights the Tensorlab implementation that will be used in the experimental Section VII. In general, the IVA framework can exploit not only the diversity provided by the presence of multiple data sets, as explained in Sections I–II, but also the diversity among samples within the same data set; see e.g. [35] for a detailed

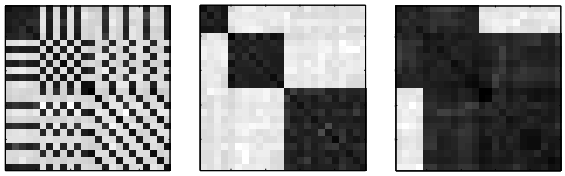
discussion of diversity in IVA. This type of diversity can occur, for example, due to time correlation or nonstationarity, and may be expressed mathematically, within a single data set, as joint diagonalization (JD) of several cumulant matrices; see, e.g., [42]. Consequently, [32], [33], [37] proposed to address the multiset scenario of IVA by factorizing several JD problems simultaneously, one for each pair (k,l). In the presence of multidimensional components ($m_i \geq 1$), this approach generalizes naturally to coupled joint block diagonalization (JBD). The JBD associated with each pair (k,l) corresponds to a tensor factorization known as rank-$(m_i^{[k]},m_i^{[l]},\cdot)$ block term decomposition (BTD) [20]. In this notation, $m_i^{[k]}$ and $m_i^{[l]}$, $i=1,\ldots,N$ indicate the row and column dimensions, respectively, of the N blocks in the (k,l)th JBD. Therefore, the Tensorlab implementation to the model in (40) amounts to a coupled rank-$(m_i,m_i,\cdot)$ BTD where $m_i^{[k]}=m_i^{[l]}=m_i$, and the third dimension of the tensor that approximates $\overline{\mathbf{X}}^{[k,l]}$ is set to one.

VII. Numerical Experiments

In this section, we validate theoretical results presented in previous sections.

A. Experimental Setup

The experimental setup is as follows. We run multiple trials for fixed $\mathbf{S}$, $\mathcal{A}$, and $\mathcal{A}_{\text{init}}$ (initial value of $\mathcal{A}$ in the algorithm), where only $\overline{\mathbf{S}}$ is drawn anew at each trial. The input data are generated such that the analysis in Section IV holds, including small-errors regime. Therefore, the theoretical value of the MSE is expected to be an accurate prediction of its empirical mean. At each trial, we compare our approach with two competing state-of-the-art methodologies. For this aim, at each trial, we test three different scenarios on the same data, as we now explain. The *first scenario* corresponds to the theoretical analysis in Section IV. Currently, there exist two algorithms that minimize (12). The Newton-based algorithm [39] converges faster than its RG [6] counterpart and is thus chosen for our numerical experiments. In this scenario, the input parameter $\mathbf{m}$ to the Newton-based algorithm [39] reflects the true block structure. In the *second scenario*, the input value of $\mathbf{m}$ to the Newton-based algorithm is set to $\mathbf{1}_{M\times 1}$, a vector of all ones. This amounts to *applying*, in a first step, the SOS-based IVA algorithm of [47] and assuming M independent sources instead of N. The latter implies that the algorithm is ignorant of the true block structure of the data and instead tries to fit it to smaller, though more numerous, blocks. Theoretically, this amounts to minimizing $C(\mathbf{A})|_{\mathbf{n}=K\mathbf{1}_M}$, in which every local minimum is also a global minimum, provided that identifiability conditions are satisfied [34]. Minimizing $C(\mathbf{A})|_{\mathbf{n}=K\mathbf{1}_M}$, however, does not imply component separation (unless the true data model is $\mathbf{m}=\mathbf{1}_M$): the M output elements are ordered arbitrarily, as explained in Section II, and a second step is required in order to cluster them into the correct N multidimensional components, as illustrated in Figures 3(a)–3(b). We denote this two-step procedure "mismodeling" [19]. In the first scenario, the clustering is implicit in the optimization via the input parameter $\mathbf{m}$. However, this comes at a price: if the permutation induced by the initialization $\mathcal{A}_{\text{init}}$ is not close enough to the permutation induced by the true $\mathcal{A}$, the algorithm may get stuck in a local minimum which is not global and fail to properly separate the blocks, in a way that cannot be compensated by a later clustering step, as illustrated in Figure 3(c). Further issues related to these two scenarios are discussed in [39].



(a) (b) (c)

Fig. 3. Example of typical clustering issues at the output of JISA algorithms, on error-free data. In this example, $\mathbf{A}=\mathbf{I}$, and the data structure is as in Figure 2. $\mathcal{A}_{\text{init}}$ is fully random, generated with $p=1$, as explained in Section VII. Figures 3(a) and 3(b) illustrate a typical output of a "mismodeling" scenario, before and after clustering, respectively. Since any local minimum is a global minimum (see explanation in Section VII-A), the correct block structure can always be reconstructed, once the permutation is found. Figure 3(c) illustrates a typical output of the algorithm when the input is the correct block structure. In this case, the random initialization often results in convergence to a local minimum that does not allow any further reconstruction of the blocks. We depict $\log 10|\cdot|$ in order to enhance small numerical features. White=zero. The color scale is arbitrary and different in each subfigure.

In the *third scenario*, the optimization uses the correct block structure but a different distance measure, as explained in Section VI. We implement (40) using Tensorlab [57], a Matlab toolbox that can straighforwardly solve coupled factorizations with a Frobenius norm. This implementation[5] is based on BTD with the third dimension set to 1, as discussed in Section VI. We optimize using `sdf_minf.m`. Due to the essential difference between the objective functions and algorithms, we take the following measures in order to allow the optimization of (40) the most favourable conditions w.r.t. (12). First, we set the stopping criteria in Tensorlab to rather small values, `TolX`=10^{-12} and `TolFun`=10^{-12}. These thresholds correspond to the relative step size and difference in objective function between every two successive iterates, respectively. Second, Tensorlab currently does not offer a straightforward positive-definite constraint on factors. Therefore, we suffice with imposing a symmetric structure on $\mathbf{\Sigma}$. This is achieved by optimizing (40) only over $k\leq l$ and attributing a double weight to the off block-diagonal factors $\mathbf{\Sigma}^{[k,l\neq k]}$. We verified that the symmetric version indeed yields better estimates of our figure of merit than leaving $\mathbf{\Sigma}$ unconstrained, as expected. Third, we initialize the algorithm with the output of the first scenario on the same samples, i.e., $\widehat{\mathcal{A}}^{\text{ML}}$, instead of $\mathcal{A}_{\text{init}}$.

The source covariance matrices are generated as $\mathbf{S}_{ii}=\text{diag}^{-\frac{1}{2}}\{\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\,\text{diag}^{-\frac{1}{2}}\{\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\}$, where $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ is the singular value decomposition (SVD) of a $Km_i\times Km_i$ matrix whose i.i.d. entries $\sim\mathcal{N}(0,1)$. The corresponding

[5]The code for the Newton and Tensorlab-based algorithms is available upon request from the authors. The Tensorlab-based code implements the general case of rank-$(m_i^{[k]},m_i^{[l]},\cdot)$ BTD with $\mathbf{A}^{[k]}$ rectangular of possibly different dimensions $\forall k$ and possibly different third dimension ("depth") for each pair (k,l).

samples are generated by right-multiplying the transpose of the Cholesky factorization of $\mathbf{S}_{ii}$ with $Km_i \times T$ i.i.d. zero-mean, unit-variance numbers, drawn from one of the following distributions: normal, or Gaussian mixture (GM) with peaks centred at ± 4. The purpose of the non-Gaussian distribution is to validate that our second-order analysis indeed holds also for non-Gaussian data. Note that right-multiplying non-Gaussian numbers with a Cholesky factorization of $\mathbf{S}_{ii}$ changes their distribution; however, it is still non-Gaussian. $\mathbf{A}^{[k]}$ is arbitrary and thus, for simplicity, fixed to $\mathbf{I}$. The stopping threshold is set to $\|\nabla C\| < 10^{-6}$, and $T = 10^4$.

In order to evaluate the MSE, our numerical validation requires not only proper reconstruction of the components but also that their ordering be the same as in the "ground truth". Trials in which this requirement is not fulfilled are easy to detect since they result in a significantly larger MSE. In this paper, we do not deal with solving these issues. Instead, and for the sake of performance analysis validation alone, we choose the initialization as $\mathbf{A}_{\text{init}}^{[k]} = p\boldsymbol{\Upsilon} + (1-p)\mathbf{I}$, where the entries of $\boldsymbol{\Upsilon}$ are $\sim \mathcal{N}(0,1)$ i.i.d. and drawn anew for each mixture k, and $p = 0.2$. This value avoids, in most cases, the need for clustering and ordering w.r.t. ground truth. In addition, for numerical stability, we choose only $\mathbf{A}_{\text{init}}^{[k]}$ whose condition number< 500. In the following simulations, trials in which mismodeling required further clustering were discarded. All other scenarios converged properly with this choice of p.

B. Numerical Results

Our results are summarized in Table I. Table I presents the normalized empirical MSE for these three scenarios, as well as its theoretical prediction, for two setups that vary in $\mathbf{m}$ and K, and thus also in $\mathbf{S}$ and $\mathcal{A}_{\text{init}}$. Each setup is tested once for Gaussian data and once for samples that are generated from numbers with a GM distribution, as explained in Section VII-A. We run 300 Monte Carlo (MC) trials; the number of trials after discarding those that did not cluster properly is indicated in the last column of Table I. The second column in Table I states the arbitrary index given to each component. The third column indicates the dimension of the ith component. The fourth column presents the theoretical prediction of the MSE per component, based only on the model parameters. Naturally, these values are not influenced by the sample distribution. The fifth column indicates the type of distribution from which the samples are generated. Columns 6–8, 9–11 and 12–14 correspond to the first, second and third scenarios, respectively. Columns 6, 9 and 12 show the averaged normalized empirical MSE per component, while columns 7, 10 and 13 provide its corresponding empirical standard deviation (std). Column 8 shows the ratio between the empirical and predicted value for the optimal scenario. Columns 11 and 14 show the ratio between the empirical MSE in the mismodeling or Frobenius norm scenarios, respectively, and the empirical MSE in the optimal case. The last row of Table I summarizes the results of certain columns. Figures 4(a) and 4(b) visualize two of the experimental configurations that are summarized in Table I: $\mathbf{m} = [3,5,4]^\top$, $K = 6$ with Gaussian data, and $\mathbf{m} = [6,5,1]^\top$, $K = 5$ with GM data, respectively. The histograms depict the distribution of the empirical MSE in MC trials (last column of Table I), as well as the empirical means and theoretical prediction.

C. Discussion of Numerical Results

The small values of the normalized $\widehat{\text{MSE}}_i$ confirm that the components have been properly separated, and quantify the quality of separation. Column 8 validates that the closed-form MSE indeed predicts the empirical mean, both for Gaussian and non-Gaussian data, as explained in Section IV-E. This also serves as an implicit validation that we are indeed in the small-errors regime. Columns 11 and 14 illustrate the potential gain in using both the correct block model and optimal norm in component separation. In particular, we observe that the gain is significant also for the estimation of one-dimensional components in the presence of multidimensional data (third component in setup #2). These observations conform with previous results on multidimensional components, previously tested only vs. the mismodeling scenario [19], [41]. When comparing the two competing methodologies, we observe that both MSE and std are generally smaller and closer to optimal in the third scenario, which uses the true block structure in the optimization but with a more relaxed norm. This observation provides further motivation for using true multidimensional methods for component and subspace separation, and developing such methods for data analysis in general.

VIII. Conclusion

In this paper, we presented a new model for simultaneous BSS of multidimensional components using SOS. We derived an ML-based component separation criterion (12). Error analysis of this criterion has led to a Newton-based algorithm [39] and to a closed-form expression for the MMSE, CRLB and FIM of the estimated parameters in the presence of real Gaussian data. For non-Gaussian data, the closed-form expression reflects the MSE when only SOS are used for the separation. We presented an alternative algebraic formulation of this criterion, as coupled matrix block-diagonalization, which can be solved by a classical model-fit approach with a Frobenius norm, and thus amounts to a new algorithm for JISA. Numerical simulations validate our theoretical analysis, and provide us with an insight on some of the assumptions that are implicit in the separation criterion, namely the choice of norm and the use of the correct block model. These preliminary results indicate that the use of a true subspace or block approach is potentially more important than the norm, and that this matter deserves to be further looked into.

The focus of this paper is on the theoretical error analysis. Therefore, numerical and practical issues such as identifying the global permutation, number or dimension of the latent sources, and choice of proper initialization in order to avoid local minima in the absence of additional information, are beyond the scope of this work.

Appendix A
Some Algebraic Properties

For ease of reference, we list some useful algebraic properties. Properties that are not proved below can be found in [53],

TABLE I

PERFORMANCE OF SECOND-ORDER JOINT INDEPENDENT SUBSPACE ANALYSIS. THEORETICALLY PREDICTED NORMALIZED MSE_i VS. EMPIRICAL, AVERAGED OVER MC TRIALS: NEWTON-BASED ALGORITHM (KLD) WITH TRUE $\mathbf{m}$ ($\widehat{\mathrm{MSE}}_i$) OR WITH $\mathbf{m}=[1,\ldots,1]^\top$ ($\widehat{\mathrm{MSE}}_i^{\mathrm{mis}}$), AND TENSORLAB (FROBENIUS NORM) WITH TRUE $\mathbf{m}$ ($\widehat{\mathrm{MSE}}_i^{\mathrm{Fro}}$). EACH MODEL SETUP IS TESTED ONCE WITH GAUSSIAN AND ONCE WITH NON-GAUSSIAN SAMPLES.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Setup	i	m_i	MSE_i	Distr.	$\widehat{\mathrm{MSE}}_i$	std	$\frac{\widehat{\mathrm{MSE}}_i}{\mathrm{MSE}_i}$	$\widehat{\mathrm{MSE}}_i^{\mathrm{mis}}$	std	$\frac{\widehat{\mathrm{MSE}}_i^{\mathrm{mis}}}{\widehat{\mathrm{MSE}}_i}$	$\widehat{\mathrm{MSE}}_i^{\mathrm{Fro}}$	std	$\frac{\widehat{\mathrm{MSE}}_i^{\mathrm{Fro}}}{\widehat{\mathrm{MSE}}_i}$	MC
	Model				KLD & correct model			KLD & mismodeling			Frobenius norm & correct model			
#1	1	3	2.35e-03	Normal	2.36e-03	3.02e-04	1.00	5.47e-03	1.22e-03	2.31	5.09e-03	6.58e-04	2.15	298
	2	5	1.63e-03		1.63e-03	1.88e-04	1.00	4.04e-03	7.17e-04	2.47	3.31e-03	3.56e-04	2.03	
	3	4	2.02e-03		2.03e-03	2.35e-04	1.00	5.06e-03	9.66e-04	2.49	3.95e-03	4.78e-04	1.94	
	1	3	2.35e-03	GM	2.38e-03	3.06e-04	1.01	5.54e-03	1.35e-03	2.32	5.02e-03	6.04e-04	2.11	268
	2	5	1.63e-03		1.65e-03	1.69e-04	1.01	4.05e-03	7.34e-04	2.45	3.30e-03	3.57e-04	2.00	
	3	4	2.02e-03		2.05e-03	2.29e-04	1.01	5.19e-03	1.12e-03	2.53	3.94e-03	4.59e-04	1.92	
#2	1	6	1.25e-03	Normal	1.26e-03	1.67e-04	1.00	3.46e-03	1.11e-03	2.75	2.64e-03	3.24e-04	2.10	274
	2	5	1.58e-03		1.59e-03	2.17e-04	1.01	4.40e-03	1.44e-03	2.76	3.35e-03	4.06e-04	2.10	
	3	1	4.99e-03		4.95e-03	1.19e-03	0.99	7.80e-03	1.92e-03	1.57	9.51e-03	2.21e-03	1.92	
	1	6	1.25e-03	GM	1.28e-03	1.57e-04	1.02	3.36e-03	8.93e-04	2.63	2.65e-03	3.56e-04	2.07	248
	2	5	1.58e-03		1.62e-03	2.01e-04	1.02	4.29e-03	1.14e-03	2.65	3.37e-03	4.62e-04	2.09	
	3	1	4.99e-03		5.06e-03	1.06e-03	1.02	7.64e-03	2.11e-03	1.51	9.62e-03	2.20e-03	1.90	
			$\ll 1$		$\ll 1$		$\cong 1$	$\ll 1$		> 1	$\ll 1$		> 1	

[58], [59].

For any matrices $\mathbf{M},\mathbf{N},\mathbf{P},\mathbf{Q}$ (with appropriate dimensions),

$$(\mathbf{N}\otimes\mathbf{M})(\mathbf{P}\otimes\mathbf{Q}) = \mathbf{NP}\otimes\mathbf{MQ} \tag{41a}$$
$$(\mathbf{N}\otimes\mathbf{M})^\top = \mathbf{N}^\top\otimes\mathbf{M}^\top \tag{41b}$$
$$\mathrm{vec}\{\mathbf{MQN}\} = (\mathbf{N}^\top\otimes\mathbf{M})\mathrm{vec}\{\mathbf{Q}\} \tag{41c}$$
$$\mathrm{tr}\{\mathbf{PQ}\} = \mathrm{tr}\{\mathbf{QP}\} \tag{41d}$$
$$\mathrm{tr}\{\mathbf{P}^\top\mathbf{Q}\} = \mathrm{vec}^\top\{\mathbf{P}\}\mathrm{vec}\{\mathbf{Q}\} \tag{41e}$$
$$\det(\mathbf{MN}) = \det(\mathbf{NM})\,. \tag{41f}$$

For any two matrices $\mathbf{M}_{M\times P}$ and $\mathbf{N}_{N\times Q}$,

$$\boldsymbol{\mathcal{T}}_{M,N}(\mathbf{N}\otimes\mathbf{M}) = (\mathbf{M}\otimes\mathbf{N})\boldsymbol{\mathcal{T}}_{P,Q}\,. \tag{42a}$$

Property A.1. Let $\mathbf{L},\mathbf{M},\mathbf{N}$ be square matrices. Then $\mathrm{tr}\{\mathbf{MLN}^\top\} = \mathrm{tr}\{(\mathbf{L}^\top\otimes\mathbf{I})\mathrm{vec}\{\mathbf{M}\}\mathrm{vec}^\top\{\mathbf{N}\}\}$. The proof follows the same steps as [41, Property A.1].

Property A.2. The first-order Taylor expansion of $(\mathbf{M}+\boldsymbol{\Delta})^{-1}$ about $\mathbf{M}$, where $\mathbf{M}$ and $\boldsymbol{\Delta}$ are square matrices, $\mathbf{M}$ and $\mathbf{M}+\boldsymbol{\Delta}$ invertible, is $(\mathbf{M}+\boldsymbol{\Delta})^{-1} = \mathbf{M}^{-1}-\mathbf{M}^{-1}\boldsymbol{\Delta}\mathbf{M}^{-1}+O(\|\boldsymbol{\Delta}\|^2)$.

APPENDIX B
ASYMPTOTIC EXPRESSION FOR THE MSE: PROOF OF (25)

Without vectorization, (24) can be equally rewritten as

$$\begin{aligned}\widehat{\mathrm{MSE}}_i^{[k]} &= \mathrm{tr}\{\overline{\mathbf{X}}^{[k,k]}(\delta\mathbf{P}_i^{[k]})^\top\delta\mathbf{P}_i^{[k]}\}\\ &= \mathrm{tr}\{\mathbf{X}^{[k,k]}(\delta\mathbf{P}_i^{[k]})^\top\delta\mathbf{P}_i^{[k]}\}\\ &\quad+\mathrm{tr}\{\delta\mathbf{X}^{[k,k]}(\delta\mathbf{P}_i^{[k]})^\top\delta\mathbf{P}_i^{[k]}\}\end{aligned} \tag{43}$$

where in the last step we have defined

$$\overline{\mathbf{X}}^{[k,k]} = \mathbf{X}^{[k,k]}+\boldsymbol{\delta}\mathbf{X}^{[k,k]}\,. \tag{44}$$

Taking expectation of (43) we obtain that

$$\begin{aligned}E\{\widehat{\mathrm{MSE}}_i^{[k]}\} &= \mathrm{tr}\{\mathbf{X}^{[k,k]}E\{(\delta\mathbf{P}_i^{[k]})^\top\delta\mathbf{P}_i^{[k]}\}\}\\ &\quad+E\{\mathrm{tr}\{\delta\mathbf{X}^{[k,k]}(\delta\mathbf{P}_i^{[k]})^\top\delta\mathbf{P}_i^{[k]}\}\}\end{aligned} \tag{45}$$

In Appendix C we show that both $\delta\mathbf{X}^{[k,k]}$ and $\delta\mathbf{P}_i^{[k]}$ are $O(\frac{1}{\sqrt{T}})$. Therefore, the second summand on the right-hand side (RHS) of (45) is (at worst) $O(\frac{1}{T^{3/2}})$, which concludes our proof.

APPENDIX C
ASYMPTOTIC PROPERTIES OF $\delta\mathbf{X}^{[k,k]}$, $\delta\mathbf{P}_i^{[k]}$, $\delta\mathbf{S}$ AND $\boldsymbol{\mathcal{E}}$

Under asymptotic conditions ($T\to\infty$), the sample covariances $\overline{\mathbf{X}}$ (defined in Section III) and $\overline{\mathbf{S}}$, and the ML estimators $\widehat{\mathbf{A}}^{[k]}$ and $\widehat{\mathbf{P}}_i^{[k]}$ (defined in Section III-B), converge, respectively, to $\mathbf{X}$, $\mathbf{S}$, $\mathbf{A}^{[k]}$ and $\mathbf{P}_i^{[k]}$, at least in probability. As for the rate of convergence, the entries of $\delta\mathbf{X}$, $\delta\mathbf{S}$, $\boldsymbol{\mathcal{E}}^{[k]}$ (26) and $\delta\mathbf{P}_i^{[k]}$ (22) are zero mean random variables with a standard deviation proportional to $1/\sqrt{T}$. For $\delta\mathbf{X}$ and $\delta\mathbf{S}$, this follows from the central limit theorem. For the ML estimation errors $\boldsymbol{\mathcal{E}}^{[k]}$ and $\delta\mathbf{P}_i^{[k]}$, this is due to sample independence. The fact that the entries of $\boldsymbol{\mathcal{E}}^{[k]}$ decrease (asymptotically) with T at the same rate as the entries of $\delta\mathbf{P}_i^{[k]}$ can also be deduced from (27). Therefore, asymptotically, the approximation $\widehat{\mathbf{P}}_i^{[k]}\cong\mathbf{P}_i^{[k]}$ holds.

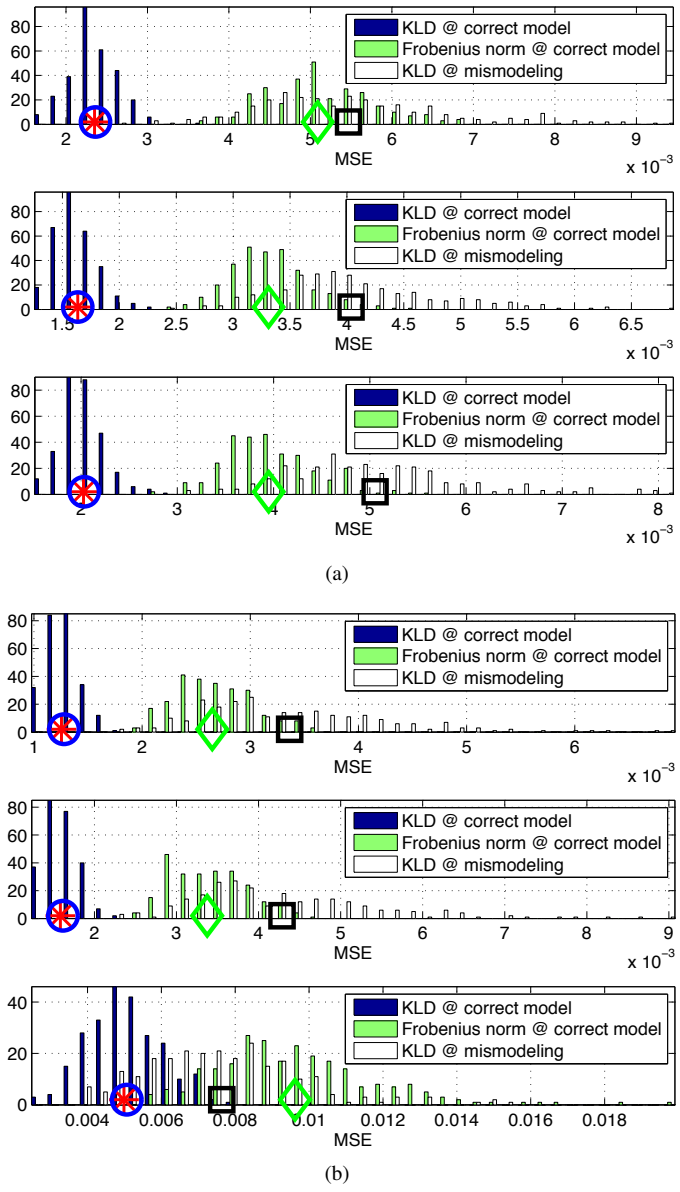
APPENDIX D
INVARIANCE OF THE ESTIMATING EQUATIONS TO SCALE AMBIGUITY OR EQUIVALENCE CLASS OF THE SOLUTIONS TO THE ESTIMATING EQUATIONS (18)

Given the existence of a set $\mathcal{A}$ that satisfies the estimating equations (18) and thus also achieves the minimum of the contrast function (12), we now discuss its equivalence class. That is, the subspace of matrices it generates that also satisfy (18) and minimize (12). Using the fact that the $\mathrm{bdiag}_{\mathbf{b}}\{\cdot\}$ operator commutes with any $\boldsymbol{\Lambda},\boldsymbol{\Lambda}'\in\mathcal{B}_{\mathbf{b}}$ in the following manner,

$$\mathrm{bdiag}_{\mathbf{b}}\{\boldsymbol{\Lambda}\mathbf{M}\boldsymbol{\Lambda}'\} = \boldsymbol{\Lambda}\,\mathrm{bdiag}_{\mathbf{b}}\{\mathbf{M}\}\boldsymbol{\Lambda}'\,,\ \forall\mathbf{M}\in\mathbb{R}^{M\times M}\,, \tag{46}$$

and given

$$\boldsymbol{\Phi}\left(\oplus_{k=1}^K\left(\oplus_{i=1}^N\mathbf{Z}_{ii}^{[k]}\right)\right) = \left(\oplus_{i=1}^N\left(\oplus_{k=1}^K\mathbf{Z}_{ii}^{[k]}\right)\right)\boldsymbol{\Phi}\,, \tag{47}$$



(a)

(b)

Fig. 4. Component separation using JISA. Histogram of normalized empirical MSE. Symbols ◯, ◊, and □ denote the empirical mean of the MSE of the depicted MC trials for KLD, Frobenius norm, and KLD mismodeling, respectively. Symbol $*$ denotes the predicted MSE for KLD. See Table I for corresponding numerical values. (a) Subplots correspond to components with dimensions 3, 5 and 4, respectively; $K = 6$, $MC = 298$, Gaussian data. (b) Subplots correspond to components with dimensions 6, 5 and 1, respectively; $K = 5$, $MC = 248$, Gaussian mixture (GM) data.

(9), (12) (41d) and (41f), we obtain that $C(\mathbf{A}) = C(\mathbf{A}\mathbf{\Lambda}'')$, where $\mathbf{\Lambda}'' \in \mathcal{B}_{\mathbf{1}_K \otimes \mathbf{m}}$ coincides with the scale ambiguity of the model (Section II). Similar steps show that the estimating equations (18) are invariant to $\mathbf{A} \to \mathbf{A}\mathbf{\Lambda}''$.

Appendix E
Derivation of (27)

In this section, we obtain (27). The proof follows steps similar to those in [19, Appendix B]. Substituting (26) in (6),

$$\begin{aligned}\widehat{\mathbf{P}}_i^{[k]} &= \mathbf{A}^{[k]}(\mathbf{I}+\boldsymbol{\mathcal{E}}^{[k]})^{-1}\mathbf{\Lambda}^{[k]}\mathbf{E}_i\mathbf{E}_i^\top\mathbf{\Lambda}^{-[k]}(\mathbf{I}+\boldsymbol{\mathcal{E}}^{[k]})\mathbf{A}^{-[k]} \\ &= \mathbf{A}^{[k]}(\mathbf{I}+\boldsymbol{\mathcal{E}}^{[k]})^{-1}\mathbf{E}_i\mathbf{E}_i^\top(\mathbf{I}+\boldsymbol{\mathcal{E}}^{[k]})\mathbf{A}^{-[k]} \\ &= \mathbf{P}_i^{[k]} + \mathbf{A}^{[k]}(\mathbf{E}_i\mathbf{E}_i^\top\boldsymbol{\mathcal{E}}^{[k]} - \boldsymbol{\mathcal{E}}^{[k]}\mathbf{E}_i\mathbf{E}_i^\top)\mathbf{A}^{-[k]} \\ &\quad + \text{higher-order terms in } \boldsymbol{\mathcal{E}}^{[k]}\,. \end{aligned} \tag{48}$$

The first equality in (48) follows from $\mathbf{A}_i^{[k]} = \mathbf{A}^{[k]}\mathbf{E}_i$ and $\mathbf{B}_i^{[k]} = \mathbf{E}_i^\top\mathbf{B}^{[k]}$, where $\mathbf{E}_i$ was defined in Section IV-B. The second equality is due to the fact that $\mathbf{\Lambda}^{[k]} \in \mathcal{B}_{\mathbf{m}}$. The third equality uses Property A.2 in Appendix A followed by $\mathbf{P}_i^{[k]} = \mathbf{A}^{[k]}\mathbf{E}_i\mathbf{E}_i^\top\mathbf{A}^{-[k]}$, which is due to (6). The last transition, from (48) to (27), follows from $\boldsymbol{\mathcal{E}}^{[k]}\mathbf{E}_i = \sum_{j=1}^N \mathbf{E}_j\boldsymbol{\mathcal{E}}_{ji}^{[k]}$ and $\mathbf{E}_i^\top\boldsymbol{\mathcal{E}}^{[k]} = \sum_{j=1}^N \boldsymbol{\mathcal{E}}_{ij}^{[k]}\mathbf{E}_j^\top$.

Appendix F
First-Order Expansion of the Estimating Equations

In this appendix, we show how a first-order expansion of the estimating equations (18) leads to the linear relation (28) between the error terms $\boldsymbol{\mathcal{E}}$ and the sample covariance matrix $\overline{\mathbf{S}}$. We begin by rewriting (18) with $\widehat{\mathbf{A}}$ instead of $\mathbf{A}$, emphasizing the fact that solutions of (18) are estimates of $\mathbf{A}$,

$$\begin{aligned}\text{bdiag}_{\mathbf{k}}\{\mathbf{\Phi}^\top\,\text{bdiag}_{\mathbf{n}}^{-1}\{\mathbf{\Phi}\widehat{\mathbf{A}}^{-1}\overline{\mathbf{X}}\widehat{\mathbf{A}}^{-\top}\mathbf{\Phi}^\top\}\mathbf{\Phi}\widehat{\mathbf{A}}^{-1}\overline{\mathbf{X}}\widehat{\mathbf{A}}^{-\top}\} \\ = \mathbf{I}\,.\end{aligned} \tag{49}$$

The link between $\boldsymbol{\mathcal{E}}$ and $\widehat{\mathbf{A}}$ is given in (26). It follows from Appendix D that the estimating equations (49) are invariant to right-multiplication of $\widehat{\mathbf{A}}$ by any matrix in $\mathcal{B}_{\mathbf{1}_K\otimes\mathbf{m}}$. Therefore, from now on, we omit the scale ambiguity term of (26). It follows from (4), (26) and the above arguments that (49) can be rewritten as

$$\begin{aligned}\text{bdiag}_{\mathbf{k}}\{\mathbf{\Phi}^\top\,\text{bdiag}_{\mathbf{n}}^{-1}\{\mathbf{\Phi}(\mathbf{I}+\boldsymbol{\mathcal{E}})\overline{\mathbf{S}}(\mathbf{I}+\boldsymbol{\mathcal{E}})^\top\mathbf{\Phi}^\top\}\mathbf{\Phi} \\ (\mathbf{I}+\boldsymbol{\mathcal{E}})\overline{\mathbf{S}}(\mathbf{I}+\boldsymbol{\mathcal{E}})^\top\} = \mathbf{I}\,.\end{aligned} \tag{50}$$

Given the factorization $\overline{\mathbf{S}} = \mathbf{S} + \delta\mathbf{S}$, one has

$$\begin{aligned}(\mathbf{I}+\boldsymbol{\mathcal{E}})\overline{\mathbf{S}}(\mathbf{I}+\boldsymbol{\mathcal{E}})^\top &= \overline{\mathbf{S}} + \boldsymbol{\mathcal{E}}\mathbf{S} + \mathbf{S}\boldsymbol{\mathcal{E}}^\top + O(\tfrac{1}{T}) \\ &= \mathbf{S} + O(\tfrac{1}{\sqrt{T}})\,,\end{aligned} \tag{51}$$

which is due to the fact that both $\boldsymbol{\mathcal{E}}$ and $\delta\mathbf{S}$ are $O(\frac{1}{\sqrt{T}})$, as explained in Appendix C. Left- and right-multiplying (51) by $\mathbf{\Phi}$ and $\mathbf{\Phi}^\top$, respectively, applying $\text{bdiag}_{\mathbf{n}}^{-1}\{\cdot\}$ and then Property A.2 in Appendix A, one obtains

$$\text{bdiag}_{\mathbf{n}}^{-1}\{\mathbf{\Phi}(\mathbf{I}+\boldsymbol{\mathcal{E}})\overline{\mathbf{S}}(\mathbf{I}+\boldsymbol{\mathcal{E}})^\top\mathbf{\Phi}^\top\} = \widetilde{\mathbf{S}}^{-1} + \widetilde{\mathbf{\Theta}} \tag{52}$$

where the entries of $\widetilde{\mathbf{\Theta}} \in \mathcal{B}_{\mathbf{n}}$ are $O(\frac{1}{\sqrt{T}})$. The term within $\text{bdiag}_{\mathbf{k}}\{\cdot\}$ in (50) can now be rewritten as

$$\begin{aligned}(\mathbf{S}^{-1}+\mathbf{\Theta})(\mathbf{I}+\boldsymbol{\mathcal{E}})\overline{\mathbf{S}}(\mathbf{I}+\boldsymbol{\mathcal{E}})^\top \\ = \mathbf{S}^{-1}\overline{\mathbf{S}} + \mathbf{S}^{-1}\boldsymbol{\mathcal{E}}\mathbf{S} + \boldsymbol{\mathcal{E}}^\top + \mathbf{\Theta}\mathbf{S} + O(\tfrac{1}{T})\end{aligned} \tag{53}$$

where $\mathbf{\Theta} = \mathbf{\Phi}^\top\widetilde{\mathbf{\Theta}}\mathbf{\Phi}$ and

$$\mathbf{\Phi}^\top(\widetilde{\mathbf{S}}^{-1} + \widetilde{\mathbf{\Theta}})\mathbf{\Phi} = \mathbf{S}^{-1} + \mathbf{\Theta}\,. \tag{54}$$

Using (53), the estimating equations can now be rewritten as

$$\begin{aligned}\text{bdiag}_{\mathbf{k}}\{\mathbf{S}^{-1}\overline{\mathbf{S}}\} + \text{bdiag}_{\mathbf{k}}\{\mathbf{S}^{-1}\boldsymbol{\mathcal{E}}\mathbf{S}\} + \boldsymbol{\mathcal{E}}^\top \\ + \text{bdiag}_{\mathbf{k}}\{\mathbf{\Theta}\mathbf{S}\} + \text{bdiag}_{\mathbf{k}}\{O(\tfrac{1}{T})\} = \mathbf{I}\end{aligned} \tag{55}$$

since $\boldsymbol{\mathcal{E}} \in \mathcal{B}_{\mathbf{k}}$.

It is clear that entries outside $\mathrm{bdiag}_{\mathbf{k}}\{\cdot\}$ do not yield any constraints on $\boldsymbol{\mathcal{E}}$. It can be further verified that entries on the main diagonal of (55) with block-pattern $\mathbf{1}_K \otimes \mathbf{m}$ are identical on both sides of (55) and thus do not have any effect. It follows that the only non-trivial terms in (49) can be written as a set of $KN(N-1)$ equations

$$[[\mathbf{S}^{-1}\overline{\mathbf{S}}]_{kk}]_{ij} + [[\mathbf{S}^{-1}\boldsymbol{\mathcal{E}}\mathbf{S}]_{kk}]_{ij} + (\boldsymbol{\mathcal{E}}_{ji}^{[k]})^\top + [[\boldsymbol{\Theta}\mathbf{S}]_{kk}]_{ij} + O(\tfrac{1}{T}) = \mathbf{0}_{m_i \times m_j}\ ,\ i \neq j \quad (56)$$

where blocks indexed by k follow block-pattern $\mathbf{k}$, those indexed by i and j follow $\mathbf{m}$, and $[[\boldsymbol{\mathcal{E}}^\top]_{kk}]_{ij} = (\boldsymbol{\mathcal{E}}_{ji}^{[k]})^\top$.

The next step is to simplify the summands in (56). For the first summand,

$$\begin{aligned}[[\mathbf{S}^{-1}\overline{\mathbf{S}}]_{kk}]_{ij} &= \mathbf{E}_i^\top \mathbf{J}_k^\top \boldsymbol{\Phi}^\top \widetilde{\mathbf{S}}^{-1} \widetilde{\overline{\mathbf{S}}} \boldsymbol{\Phi} \mathbf{J}_k \mathbf{E}_j \\ &= \mathbf{U}_k^\top \mathbf{Y}_i^\top \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \widetilde{\mathbf{S}}^{-1} \widetilde{\overline{\mathbf{S}}} \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \mathbf{Y}_j \mathbf{U}_k \\ &= \mathbf{U}_k^\top \mathbf{S}_{ii}^{-1} \overline{\mathbf{S}}_{ij} \mathbf{U}_k = [\mathbf{S}_{ii}^{-1}\overline{\mathbf{S}}_{ij}]_{kk} \end{aligned} \quad (57)$$

as we now explain. The first step uses (7) such that $\mathbf{S}^{-1}\overline{\mathbf{S}} = \boldsymbol{\Phi}^\top \widetilde{\mathbf{S}}^{-1} \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \widetilde{\overline{\mathbf{S}}} \boldsymbol{\Phi} = \boldsymbol{\Phi}^\top \widetilde{\mathbf{S}}^{-1} \widetilde{\overline{\mathbf{S}}} \boldsymbol{\Phi}$. $\mathbf{J}_k$ and $\mathbf{E}_i$ were defined in Sections III-B and IV-B, respectively. In the second step of (57) we employ

$$\mathbf{J}_k \mathbf{E}_j = \boldsymbol{\Phi}^\top \mathbf{Y}_j \mathbf{U}_k \quad (58)$$

where $\mathbf{U}_k$ is $n_i \times m_i$ and $\mathbf{Y}_i$ is $L \times n_i$ such that $\mathbf{I}_{n_i} = [\mathbf{U}_1|\cdots|\mathbf{U}_K]$ and $\mathbf{I}_L = [\mathbf{Y}_1|\cdots|\mathbf{Y}_N]$. The third step uses $[\widetilde{\mathbf{S}}^{-1}\widetilde{\overline{\mathbf{S}}}]_{ij} = \mathbf{S}_{ii}^{-1}\overline{\mathbf{S}}_{ij}$, which follows from $\widetilde{\mathbf{S}} \in \mathcal{B}_{\mathbf{n}}$. For the second summand in (56), we begin by writing explicitly the term within $[\cdot]_{ij}$,

$$\begin{aligned}[\mathbf{S}^{-1}\boldsymbol{\mathcal{E}}\mathbf{S}]_{kk} &= \sum_{l=1}^K \mathbf{J}_k^\top \mathbf{S}^{-1} \mathbf{J}_l \boldsymbol{\mathcal{E}}^{[l]} \mathbf{J}_l^\top \mathbf{S} \mathbf{J}_k \\ &= \sum_{l=1}^K [\mathbf{S}^{-1}]_{kl} \boldsymbol{\mathcal{E}}^{[l]} \mathbf{S}^{[l,k]} \end{aligned} \quad (59)$$

where $\boldsymbol{\mathcal{E}} \triangleq \sum_{l=1}^K \mathbf{J}_l \boldsymbol{\mathcal{E}}^{[l]} \mathbf{J}_l^\top$. It follows from (7) that $\mathbf{S}^{-1}$ has the same zero-pattern as $\mathbf{S}$ such that $[\mathbf{S}^{-1}]_{kl} \in \mathcal{B}_{\mathbf{m}}$ and $[[\mathbf{S}^{-1}]_{kl}]_{ii} = [\mathbf{S}_{ii}^{-1}]_{kl}$. Hence, the (i,j)th block of the lth summand on the RHS of (59) can be factorized as

$$[[\mathbf{S}^{-1}]_{kl}\boldsymbol{\mathcal{E}}^{[l]}\mathbf{S}^{[l,k]}]_{ij} = [[\mathbf{S}^{-1}]_{kl}]_{ii}\boldsymbol{\mathcal{E}}_{ij}^{[l]}\mathbf{S}_{jj}^{[l,k]} = [\mathbf{S}_{ii}^{-1}]_{kl}\boldsymbol{\mathcal{E}}_{ij}^{[l]}\mathbf{S}_{jj}^{[l,k]}\ .$$

This concludes the derivation of the second summand in (56). The third summand in (56) remains unchanged. We conclude the derivation of the first equation in (28) by showing that $[[\boldsymbol{\Theta}\mathbf{S}]_{kk}]_{ij} = \mathbf{0}_{m_i \times m_j}$. This follows from noting that $\boldsymbol{\Theta}\mathbf{S} = \boldsymbol{\Phi}^\top \widetilde{\boldsymbol{\Theta}} \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \widetilde{\mathbf{S}} \boldsymbol{\Phi} = \boldsymbol{\Phi}^\top \widetilde{\boldsymbol{\Theta}} \widetilde{\mathbf{S}} \boldsymbol{\Phi}$ has the same zero-pattern as $\mathbf{S}$ and by definition, $[[\mathbf{S}]_{kk}]_{ij} = \mathbf{0}_{m_i \times m_j}$. The second equation in (28) is obtained by exchanging i and j.

Appendix G
Closed-Form Expression for $\mathrm{Cov}(\mathbf{g})$

In this Appendix we derive a closed-form expression for the covariance of the gradient vectors $\mathbf{g}_{ij}$, defined in (31). By the assumptions in Section III, these gradients have zero mean. We now show that

$$E\{\mathbf{g}_{ij}\mathbf{g}_{mn}^\top\} = \begin{cases} \frac{1}{T}(\mathbf{S}_{jj} \boxtimes \mathbf{S}_{ii}^{-1}) & (m,n) = (i,j) \\ \frac{1}{T}(\mathbf{I}_K \otimes \boldsymbol{\mathcal{T}}_{m_j,m_i}) & (m,n) = (j,i) \\ \mathbf{0} & \text{o.w.} \end{cases} \quad (60)$$

The building blocks of (60) are terms of the type

$$\begin{aligned}[E\{\mathbf{g}_{ij}\mathbf{g}_{mn}^\top\}]_{kl} = \sum_{\alpha=1}^K \sum_{\beta=1}^K (\mathbf{I} \otimes [\mathbf{S}_{ii}^{-1}]_{k\alpha}) \\ E\{\mathrm{vec}\{[\overline{\mathbf{S}}_{ij}]_{\alpha k}\}\mathrm{vec}^\top\{[\overline{\mathbf{S}}_{mn}]_{\beta l}\}\}(\mathbf{I} \otimes [\mathbf{S}_{mm}^{-1}]_{l\beta})^\top\ , \end{aligned} \quad (61)$$

that relate the covariance of the gradients to the sample covariance, for any $i,j,m,n \in \{1,\ldots,N\}$ and $k,l \in \{1,\ldots,K\}$, as we now explain. In (61) we reformulated the kth $m_i m_j \times 1$ term of $\mathbf{g}_{ij}$ as

$$\mathrm{vec}\{[\mathbf{S}_{ii}^{-1}\overline{\mathbf{S}}_{ij}]_{kk}\} = \sum_{\alpha=1}^K (\mathbf{I} \otimes [\mathbf{S}_{ii}^{-1}]_{k\alpha})\mathrm{vec}\{[\overline{\mathbf{S}}_{ij}]_{\alpha k}\}\ , \quad (62)$$

which follows from applying $\mathrm{vec}\{\cdot\}$ to

$$[\mathbf{S}_{ii}^{-1}\overline{\mathbf{S}}_{ij}]_{kk} = \sum_{\alpha=1}^K [\mathbf{S}_{ii}^{-1}]_{k\alpha}[\overline{\mathbf{S}}_{ij}]_{\alpha k} \quad (63)$$

and then (41c) in order to separate the stochastic and the deterministic terms. The equality in (63) follows from $\mathbf{S}_{ii} \in \mathcal{B}_{\mathbf{1}_K \otimes m_i}$. Using the explicit form

$$[\overline{\mathbf{S}}_{ij}]_{\alpha k} = \frac{1}{T}\sum_{t=1}^T \mathbf{s}_i^{[\alpha]}(t)\mathbf{s}_j^{[k]\top}(t) \quad (64)$$

and following steps similar to those in [41, Appendix D], one obtains

$$[E\{\mathbf{g}_{ij}\mathbf{g}_{mn}^\top\}]_{kl} = \begin{cases} \frac{1}{T}(\mathbf{S}_{jj}^{[k,l]} \otimes [\mathbf{S}_{ii}^{-1}]_{kl}) & (m,n) = (i,j) \\ \frac{1}{T}\boldsymbol{\mathcal{T}}_{m_j,m_i}\delta_{kl} & (m,n) = (j,i) \\ \mathbf{0} & \text{o.w.} \end{cases}$$

which is the block-wise form of (60). As in [41, Appendix D], the derivation is based on $E\{\mathbf{s}_i(t)\mathbf{s}_{j\neq i}^\top(r)\} = \mathbf{0}\ \forall r,t$, $E\{\mathbf{s}_i(t)\} = \mathbf{0}\ \forall i,t$ and sample decorrelation (in accordance with Section III), without resorting to any further assumptions on the distributions of the samples.

Acknowledgement

The authors would like to thank the anonymous reviewers for their careful reading and valuable remarks.

References

[1] P. Comon, "Supervised classification, a probabilistic approach," in *Proc. ESANN*, Brussels, Belgium, Apr. 1995, pp. 111–128.

[2] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Fetal electrocardiogram extraction by source subspace separation," in *Proc. IEEE SP/ATHOS Workshop on HOS*, Girona, Spain, Jun. 1995, pp. 134–138.

[3] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. ICASSP*, vol. 4, Seattle, WA, May 1998, pp. 1941–1944.

[4] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Independent Component Analysis and Blind Signal Separation*, ser. LNCS, vol. 3889. Springer Berlin Heidelberg, 2006, pp. 165–172.

[5] Y.-O. Li, T. Adalı, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3918–3929, Oct. 2009.

[6] D. Lahat and C. Jutten, "Joint blind source separation of multidimensional components: Model and algorithm," in *Proc. EUSIPCO*, Lisbon, Portugal, Sep. 2014, pp. 1417–1421.

[7] F. J. Theis, "Towards a general independent subspace analysis," in *Proc. NIPS*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 1361–1368.

[8] J. A. Palmer and S. Makeig, "Contrast functions for independent subspace analysis," in *Latent Variable Analysis and Signal Separation*, ser. LNCS, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds., vol. 7191. Springer, 2012, pp. 115–122.

[9] Z. Szabó, B. Póczos, and A. Lőrincz, "Separation theorem for independent subspace analysis and its consequences," *Pattern Recognition*, vol. 45, no. 4, pp. 1782–1791, Apr. 2012.

[10] P. Tichavský, A. Yeredor, and Z. Koldovský, "On computation of approximate joint block-diagonalization using ordinary AJD," in *Latent Variable Analysis and Signal Separation*, ser. LNCS, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds., vol. 7191. Springer, 2012, pp. 163–171.

[11] A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," *Neural Comput.*, vol. 12, no. 7, pp. 1705–1720, Jul. 2000.

[12] L. De Lathauwer, C. Févotte, B. De Moor, and J. Vandewalle, "Jacobi algorithm for joint block diagonalization in blind identification," in *Proc. 23rd Symp. Inf. Theory in the Benelux*, Louvain-la-Neuve, Belgium, May 29–31 2002, pp. 155–162.

[13] A. Hyvärinen and U. Köster, "FastISA: A fast fixed-point algorithm for independent subspace analysis," in *Proc. ESANN*, Bruges (Belgium), Apr. 2006, pp. 371–376.

[14] L. De Lathauwer and D. Nion, "Decompositions of a higher-order tensor in block terms. Part III: Alternating least squares algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1067–1083, 2008.

[15] P. Tichavský and Z. Koldovský, "Algorithms for nonorthogonal approximate joint block-diagonalization," in *Proc. EUSIPCO*, Bucharest, Romania, Aug. 2012, pp. 2094–2098.

[16] D. Lahat, J.-F. Cardoso, and H. Messer, "Joint block diagonalization algorithms for optimal separation of multidimensional components," in *Latent Variable Analysis and Signal Separation*, ser. LNCS, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds., vol. 7191. Springer, 2012, pp. 155–162.

[17] L. Sorber, M. Van Barel, and L. De Lathauwer, "Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-$(L_r, L_r, 1)$ terms, and a new generalization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 695–720, 2013.

[18] M. Sørensen, I. Domanov, and L. De Lathauwer, "Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(L_{r,n}, L_{r,n}, 1)$ terms—part II: Algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 3, pp. 1015–1045, Jul. 2015.

[19] D. Lahat, J.-F. Cardoso, and H. Messer, "Blind separation of multidimensional components via subspace decomposition: Performance analysis," *IEEE Trans. Signal Process.*, vol. 62, no. 11, pp. 2894–2905, Jun. 2014.

[20] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms. Part II: Definitions and uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1033–1066, 2008.

[21] D. Lahat, J.-F. Cardoso, and H. Messer, "Identifiability of second-order multidimensional ICA," in *Proc. EUSIPCO*, Bucharest, Romania, Aug. 2012, pp. 1875–1879.

[22] H. W. Gutch and F. J. Theis, "Uniqueness of linear factorizations into independent subspaces," *J. Multivar. Anal.*, vol. 112, pp. 48–62, Nov. 2012.

[23] M. Sørensen and L. De Lathauwer, "Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(L_{r,n}, L_{r,n}, 1)$ terms—part I: Uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 2, pp. 496–522, Apr. 2015.

[24] D. Lahat and C. Jutten, "A generalization to Schur's lemma with an application to joint independent subspace analysis," GIPSA-Lab, Grenoble, France, Tech. Rep. hal-01247899, Dec. 2015.

[25] S. Ma, N. M. Correa, X.-L. Li, T. Eichele, V. D. Calhoun, and T. Adalı, "Automatic identification of functional clusters in fMRI data using spatial dependence," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 12, pp. 3406–3417, Dec. 2011.

[26] B. Hunyadi, D. Camps, L. Sorber, W. Van Paesschen, M. De Vos, S. Van Huffel, and L. De Lathauwer, "Block term decomposition for modelling epileptic seizures," *EURASIP JASP*, vol. 2014, no. 14-04, p. 139, 2014.

[27] H. Bousbia-Salah, A. Belouchrani, and K. Abed-Meraim, "Blind separation of convolutive mixtures using joint block diagonalization," *Proc. ISSPA 2001*, vol. 1, pp. 13–16, Aug. 2001.

[28] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[29] J.-F. Cardoso, M. Le Jeune, J. Delabrouille, M. Betoule, and G. Patanchon, "Component separation with flexible models – application to multichannel astrophysical observations," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 5, pp. 735–746, Oct. 2008.

[30] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, Dec. 1936.

[31] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[32] X.-L. Li, T. Adalı, and M. Anderson, "Joint blind source separation by generalized joint diagonalization of cumulant matrices," *Signal Process.*, vol. 91, no. 10, pp. 2314–2322, Oct. 2011.

[33] M. Congedo, R. Phlypo, and J. Chatel-Goldman, "Orthogonal and non-orthogonal joint blind source separation in the least-squares sense," in *Proc. EUSIPCO*, Bucharest, Romania, Aug. 2012, pp. 1885–1889.

[34] M. Anderson, T. Adalı, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1672–1683, Apr. 2012.

[35] T. Adalı, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Process. Mag.*, pp. 18–33, May 2014.

[36] T. W. Anderson, *An introduction to multivariate statistical analysis*. John Wiley & Sons, 1958.

[37] J. Chatel-Goldman, M. Congedo, and R. Phlypo, "Joint BSS as a natural analysis framework for EEG-hyperscanning." in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 1212–1216.

[38] S. Ma, V. D. Calhoun, R. Phlypo, and T. Adalı, "Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis," *NeuroImage*, vol. 90, pp. 196–206, Apr. 2014.

[39] D. Lahat and C. Jutten, "Joint independent subspace analysis: A quasi-Newton algorithm," in *Latent Variable Analysis and Signal Separation*, ser. LNCS, vol. 9237. Springer International Publishing Switzerland, 2015, pp. 111–118, in Proc. LVA/ICA 2015, Liberec, Czech Republic, Aug.

[40] R. F. Silva, S. Plis, T. Adalı, and V. D. Calhoun, "Multidataset independent subspace analysis extends independent vector analysis," in *Proc. ICIP*, Paris, France, Oct. 2014, pp. 2864–2868.

[41] D. Lahat, J.-F. Cardoso, and H. Messer, "Second-order multidimensional ICA: Performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4598–4610, Sep. 2012.

[42] J.-F. Cardoso, "The three easy routes to independent component analysis; contrasts and geometry," in *Proc. ICA*, San Diego, CA, USA, Dec. 2001, pp. 1–6.

[43] D. Lahat, T. Adalı, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[44] R. A. Horn and R. Mathias, "Block-matrix generalizations of Schur's basic theorems on Hadamard products," *Linear Algebra and its Applications*, vol. 172, pp. 337–346, Jul. 1992.

[45] S. Liu, "Matrix results on the Khatri-Rao and Tracy-Singh products," *Linear Algebra and its Applications*, vol. 289, no. 1–3, pp. 267–277, Sep. 1999.

[46] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[47] J. Vía, M. Anderson, X.-L. Li, and T. Adalı, "A maximum likelihood approach for independent vector analysis of Gaussian data sets," in *Proc. MLSP*, Beijing, China, Sep. 2011.

[48] P. Comon, "Independent component analysis," in *Proc. Int. Signal Process. Workshop on HOS*, Chamrousse, France, Jul. 1991, pp. 111–120, keynote address. Republished in *HOS*, J.-L. Lacoume ed., Elsevier, 1992, pp. 29–38.

[49] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.

[50] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[51] D.-T. Pham and P. Garat, "Blind separation of mixtures of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. Signal Process.*, vol. 45, no. 7, pp. 1712–1725, Jul. 1997.

[52] D.-T. Pham, "Joint approximate diagonalization of positive definite hermitian matrices," *SIAM J. Matrix Anal. Appl.*, vol. 22, no. 4, pp. 1136–1152, 2001.

[53] J. R. Magnus and H. Neudecker, "The commutation matrix: Some properties and applications," *Ann. Statist.*, vol. 7, no. 2, pp. 381–394, Mar. 1979.

[54] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4399–4410, Sep. 2014.

[55] G. Chabriel, M. Kleinsteuber, E. Moreau, H. Shen, P. Tichavský, and A. Yeredor, "Joint matrices decompositions and blind source separation: A survey of methods, identification, and applications," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 34–43, 2014.

[56] L. Sorber, M. Van Barel, and L. De Lathauwer, "Structured data fusion," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 4, pp. 586–600, Jun. 2015.

[57] ——, "Tensorlab v2.0," Jan. 2014. [Online]. Available: http://www.tensorlab.net/

[58] A. Graham, *Kronecker Products and Matrix Calculus with Applications*, ser. Mathematics and its Applications. Chichester, West Sussex, England: Ellis Horwood Limited, 1981.

[59] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Nov. 2012, version 20121115. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?3274