



HAL
open science

Second-Order Multidimensional ICA: Performance Analysis

Dana Lahat, Jean-François Cardoso, Hagit Messer

► **To cite this version:**

Dana Lahat, Jean-François Cardoso, Hagit Messer. Second-Order Multidimensional ICA: Performance Analysis. IEEE Transactions on Signal Processing, 2012, 60 (9), pp.4598-4610. 10.1109/TSP.2012.2199985 . hal-02286750

HAL Id: hal-02286750

<https://telecom-paris.hal.science/hal-02286750v1>

Submitted on 13 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Second-Order Multidimensional ICA: Performance Analysis

Dana Lahat, Jean-François Cardoso, and Hagit Messer, *Fellow, IEEE*

Abstract—Independent component analysis (ICA) and blind source separation (BSS) deal with extracting a number of mutually independent elements from a set of observed linear mixtures. Motivated by various applications, this paper considers a more general and more flexible model: the sources can be partitioned into groups exhibiting dependence within a given group but independence between two different groups. We argue that this is tantamount to considering *multidimensional* components as opposed to the standard ICA case which is restricted to one-dimensional components. The core of the paper is devoted to the statistical analysis of the blind separation of multidimensional components based on second-order statistics, in a piecewise-stationary model. We develop the likelihood and the associated estimating equations for the Gaussian case. We obtain closed-form expressions for the Fisher information matrix and the Cramér–Rao bound of the de-mixing parameters, as well as the mean-square error (MSE) of the component estimates. The derived MSE is valid also for non-Gaussian data. Our analysis is verified through numerical experiments, and its performance is compared to classical ICA in various dependence scenarios, quantifying the gain in the accuracy of component recovery in presence of multidimensional components.

Index Terms—Blind source separation, joint block diagonalization, multidimensional independent component analysis, performance analysis, second-order methods.

I. INTRODUCTION

IN their most basic setting, independent component analysis (ICA) and blind source separation (BSS) aim at extracting m mutually independent elements from m observed mixtures. The model is of T observations of an $m \times 1$ vector $\mathbf{x}(t)$, modeled as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \quad 1 \leq t \leq T \quad (1)$$

where \mathbf{A} is an $m \times m$ full-rank matrix and $\mathbf{s}(t)$ is a vector of independent *sources*. A natural extension of practical interest is to assume that the m sources can be partitioned into $n \leq m$ groups with the sources of different groups being statistically independent while the sources in the same group are not independent and cannot be made independent by any linear transform on $\mathbf{s}(t)$. In the following, we use the term *dependent sources* to

indicate such a source model. As we shall see shortly, the multiplicative model (1) of dependent sources, which suffers from inherent indeterminacies, is equivalent to the additive model [1]

$$\mathbf{x}(t) = \sum_{i=1}^n \mathbf{x}_i(t) \quad (2)$$

of *multidimensional components*, in which these indeterminacies are avoided.

Multidimensional data may occur due to various complex relations within the dependent elements. The dimension of a dependent group may not always reflect the actual number of its underlying elements. As a result, in multidimensional models, there is not always a physically meaningful interpretation to separating the multidimensional components back into single-dimensional elements. Various phenomena can yield data that is multidimensional in nature. For example, in the analysis of MEG stimuli [2], energies of evoked signals may become statistically dependent due to similar activation and termination times. In addition, a single isolated neural source is of little diagnostic value, and consequently the number of detected sources may be reduced through a clustering procedure, based on spatial topography [3]. In natural image analysis, subspaces originate from features which represent positions, orientations, spatial frequencies and phases [4]. In fetal ECG detection [1], [5], multidimensionality is explained by a certain 3D electric model of the heart, and may change during pregnancy. Finally, in the separation of astrophysical emissions, multidimensionality may reflect not only dependence between different emissions due to astrophysical processes [6], but also spatial nonstationarity, when observations are taken over a large enough patch of the sky [7], [8].

The idea of solving the dependent sources/multidimensional components problem in terms of subspace separation through ICA was first demonstrated in [5], on fetal ECG recordings. The perspective of *multidimensional ICA* (MICA), of vector-valued components whose representation is based on unambiguous projections on the sources' respective subspaces, was presented in [5], and an elaborate geometric framework was suggested in [1]. The notion of independent feature subspaces and independent subspace analysis (ISA) was introduced in [4] and [9]. In the ISA framework, independent and identically distributed (i.i.d.) non-Gaussian observations are separated into feature-invariant subspaces by a criterion which maximizes the independence between norms of projections on linear subspaces. ISA algorithms usually assume that the sources have a spherically symmetric distribution; examples are given in [4], [10]–[12]. This can be obtained in practice by a preliminary whitening stage. Whitening simplifies the optimization by restricting the de-mixing matrix to be orthogonal; this results in a smaller number of variables to adjust. However, by assuming

D. Lahat and H. Messer are with the School of Electrical Engineering, Tel Aviv University, 69978 Tel Aviv, Israel (e-mail: danal@eng.tau.ac.il; messer@eng.tau.ac.il).

J.-F. Cardoso is with LTCI, TELECOM ParisTech, 75013 Paris, France (e-mail: cardoso@tsi.enst.fr).

that the *empirical* covariance of the observations is the identity matrix, whitening ignores finite-data effects. Whitening errors which are introduced in this preliminary stage cannot be compensated for by the following rotation stage [13]. Hence, such methods are sub-optimal with respect to methods which do not preprocess the data by whitening [14], [15]. An algorithm which solves ISA through joint block diagonalization (JBD) of cumulant matrices, where JBD is performed by joint diagonalization (JD) and permutation-recovery, with a pre-whitening stage, is described in [16]. An ISA algorithm without the whitening constraint is given, for example, in [17]. A non-parametric dependence measure, whose minimization is based on one-dimensional ICA as a step in separating multidimensional channels for non-Gaussian i.i.d. pre-whitened data, is found in [18]. The above-mentioned algorithms require as an input the correct subspace dimensions. An algorithm which extends [10] by automatically detecting the subspace dimensions and clustering the data is proposed in [19]. Conditions under which the two-step procedure of ICA and then clustering is sufficient for separation are discussed in [20]; this analysis is valid for i.i.d. pre-whitened data and specific distribution types.

In the above-mentioned works, the performance of the algorithms is examined numerically [1], [4], [5], [9], [18], [19], or their convergence to the correct separation point is discussed theoretically [10]–[12], [16], [17], [20]. However, to the best of our knowledge, no complete performance analysis, in the sense of closed-form expressions for an expected figure of merit or bound, has been conducted to any of these scenarios.

Cramér–Rao lower bounds (CRLB) on the estimation error of the mixing matrix and of the source parameters in ICA have already been discussed in the literature. Optimal performance and lower bounds were analyzed, for example, for the following scenarios: system identification for non-Gaussian sources [21]; second-order static-mixture noisy ICA [22]; noise-free static-mixture ICA for non-Gaussian sources [23]–[25]; stationary parametric Gaussian sources [26] and non-stationary sources [27]. A lower bound on source separation due to whitening has been derived in [14] and [15].

In this paper, we consider a second-order based method which extends the maximum likelihood (ML) treatment of [28] and the performance analysis of [27] and [29] to the case of non-stationary Gaussian dependent sources/multidimensional components. We provide a complete small-error performance analysis, in the sense of closed-form expressions for the Fisher information matrix (FIM), CRLB and mean-square error (MSE). We do not discuss algorithms which achieve these measures; such algorithms can be found in [30]. As in [28], there is no whitening constraint. Preliminary results can be found in [31].

It should be noted that the Gaussian non-stationary analysis, discussed in this paper, is also appropriate for stationary non-white processes. As demonstrated by [32] and [33], in full analogy to the time-domain procedure, second-order ICA can also be performed in the frequency-domain on non-white processes, due to the asymptotic properties of the Fourier coefficients. In this case, the t index implies a frequency index. An application of the proposed performance analysis to the separation of astrophysical emissions, in the frequency domain, can be found in [8].

The structure of this paper is as follows. In Section II, we present the two points of view, of dependent sources versus mul-

tidimensional components. In Section III, we describe our statistical model. We derive a contrast function, whose minimization is equivalent to the ML solution for Gaussian piecewise stationary data. We show that this minimization can be obtained by joint block diagonalization (JBD) of a set of sample covariance matrices. We then derive the estimating equations, whose component-wise form forms the basis for the detailed error analysis in Section IV. The small-error analysis in Section IV is done exclusively in terms of well-defined quantities, i.e., components and projections. From the Taylor expansion of the estimating equations, we calculate the error covariance of the ML estimates, and this result is propagated to yield the total MSE in component estimation. Section V deals with identifiability of the model. Namely, the required number of stationary intervals, as well as conditions on the source covariance matrices, to guarantee existence, uniqueness and identifiability of the model. Numerical results are presented in Section VI. Numerical simulations demonstrate that our CRLB is achievable when the model assumptions hold and that our MSE expression for component reconstruction holds also for non-Gaussian data, as expected from our theoretical analysis. Furthermore, the simulations show a significant gain of the multidimensional approach over the cruder approach based on one-dimensional separation following a clustering of the dependent sources, as suggested in [34].

The following notations and conventions are used throughout this paper. Bold lowercase letters denote vectors; regular lowercase letters denote scalars; bold uppercase letters denote matrices; regular uppercase letters denote functions, operators or constants. Transposition is denoted by \dagger ; trace is denoted by $\text{tr}\{\cdot\}$; $|\mathbf{a}|^2 = \mathbf{a}^\dagger \mathbf{a}$ for any vector \mathbf{a} . $\text{Cov}(\mathbf{a}) \triangleq E\{\mathbf{a}\mathbf{a}^\dagger\}$, $\text{Cov}(\mathbf{a}, \mathbf{b}) \triangleq E\{\mathbf{a}\mathbf{b}^\dagger\}$ for any stochastic vectors \mathbf{a}, \mathbf{b} with $E\{\mathbf{a}\} = \mathbf{0}$. $\|\cdot\|^2$ denotes the Frobenius norm; δ_{ij} denotes the Kronecker delta. The Kronecker product is denoted by \otimes . $\Omega(f)$ stands for zero mean stochastic terms whose standard deviation is proportional to f , or to higher powers thereof. $O(f)$ stands for deterministic terms which are bounded above, up to a constant factor, by f .

II. DEPENDENT SOURCES VERSUS MULTIDIMENSIONAL COMPONENTS

The separation problem of interest can be stated by partitioning the source vector $\mathbf{s}(t)$ and matrix \mathbf{A} as

$$\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_n], \quad \mathbf{s}(t) = [\mathbf{s}_1^\dagger(t), \dots, \mathbf{s}_n^\dagger(t)]^\dagger \quad (3)$$

where \mathbf{A}_i , the i th column block of \mathbf{A} , has dimension $m \times m_i$, vector $\mathbf{s}_i(t)$ has dimension $m_i \times 1$ and $\sum_{i=1}^n m_i = m$. Given the pattern $\mathbf{m} \triangleq [m_1, \dots, m_n]^\dagger$ and the observations $\mathbf{x}(t)$, the problem of blind separation of dependent sources is that of finding matrices \mathbf{A}_i such that \mathbf{A} is full-rank and such that the corresponding source vectors $\mathbf{s}_1(t), \dots, \mathbf{s}_n(t)$ are *as independent as possible*. This notion is given a definite meaning in Section III where we set up a simple statistical model which, via its likelihood function, yields a quantitative measure of independence. However, before we write down a likelihood function, it is necessary to discuss the indeterminacies inherent to the blind separation of dependent sources and to explain how these indeter-

minacies lead to the alternate point of view of multidimensional ICA.

With the partition (3), the multiplicative, source-mixing model (1) can also be written as an additive model (2), where we define the i th component $\mathbf{x}_i(t)$ as the $m \times 1$ vector

$$\mathbf{x}_i(t) = \mathbf{A}_i \mathbf{s}_i(t). \quad (4)$$

In a blind context, the component vector $\mathbf{x}_i(t)$ is better defined than the source vector $\mathbf{s}_i(t)$. Indeed, for any $m_i \times m_i$ invertible matrix \mathbf{Z}_i , the pair $(\mathbf{A}_i, \mathbf{s}_i(t))$ and the pair $(\mathbf{A}_i \mathbf{Z}_i^{-1}, \mathbf{Z}_i \mathbf{s}_i(t))$ contribute the same quantity $\mathbf{x}_i(t) = \mathbf{A}_i \mathbf{Z}_i^{-1} \mathbf{Z}_i \mathbf{s}_i(t) = \mathbf{A}_i \mathbf{s}_i(t)$ to the observations. It is thus impossible to discriminate between the representation of a component $\mathbf{x}_i(t)$ by the pair $(\mathbf{A}_i, \mathbf{s}_i(t))$ and by the pair $(\mathbf{A}_i \mathbf{Z}_i^{-1}, \mathbf{Z}_i \mathbf{s}_i(t))$. Therefore, matrix \mathbf{A}_i can, at best, be blindly identified only up to right multiplication by an invertible $m_i \times m_i$ matrix. This is the familiar scale indeterminacy of standard ICA, carried over to dimension m_i , with an $m_i \times m_i$ matrix factor \mathbf{Z}_i instead of a simple scalar factor.

Since matrix \mathbf{A}_i is determined only up to a right factor \mathbf{Z}_i , only its *column space*, $\text{Span}(\mathbf{A}_i)$, can be blindly identified. It is thus useful to introduce the separating *projectors*: these are the $m \times m$ oblique projection matrices \mathbf{P}_i onto $\text{Span}(\mathbf{A}_i)$ along $\text{Span}(\mathbf{A}_j)$ for all $j \neq i$. By definition, they satisfy $\mathbf{P}_i \mathbf{A}_j = \delta_{ij} \mathbf{A}_i$, are unaffected if \mathbf{A}_i is changed into $\mathbf{A}_i \mathbf{Z}_i^{-1}$, and allow one to write

$$\mathbf{x}_i(t) = \mathbf{P}_i \mathbf{x}(t). \quad (5)$$

The notation (5) is the geometric counterpart of (4). The set of unambiguous oblique projections \mathbf{P}_i is the matrix-free counterpart of the inverse matrix $\mathbf{B} = \mathbf{A}^{-1}$.

For later reference we mention that if \mathbf{B} is partitioned into n horizontal blocks, where the i th block \mathbf{B}_i has dimension $m_i \times m$, then the i th oblique projector is given by

$$\mathbf{P}_i = \mathbf{A}_i \mathbf{B}_i. \quad (6)$$

We also define the orthogonal projection matrix $\mathbf{\Pi}_i$ onto $\text{Span}(\mathbf{A}_i)$, that is,

$$\mathbf{\Pi}_i = \mathbf{A}_i (\mathbf{A}_i^\dagger \mathbf{A}_i)^{-1} \mathbf{A}_i^\dagger = \mathbf{A}_i \mathbf{A}_i^\sharp, \quad (7)$$

denoting by $\mathbf{A}_i^\sharp = (\mathbf{A}_i^\dagger \mathbf{A}_i)^{-1} \mathbf{A}_i^\dagger$ the Moore–Penrose pseudoinverse (49) of \mathbf{A}_i . Obviously, projector $\mathbf{\Pi}_i$ is unaffected if \mathbf{A}_i is changed into $\mathbf{A}_i \mathbf{Z}_i^{-1}$.

In summary, the source separation model based on a mixing matrix is recast as a component separation model (2), where the i th component $\mathbf{x}_i(t)$ is restricted to an m_i -dimensional subspace, represented by the uniquely defined orthogonal projector $\mathbf{\Pi}_i$. The i th component is recovered via (5) using the oblique projector \mathbf{P}_i .

As a final note in this section, we would like to emphasize that the component perspective goes beyond the mere avoidance of scale indeterminacy. The various examples of multidimensional data in Section I demonstrate that in a multidimensional setup, a “mixing matrix” and “dependent sources” may not always have a physical interpretation of their own.

III. MODEL, LIKELIHOOD AND CONTRAST FUNCTION

We derive a likelihood function for the separation of dependent sources by generalizing the Gaussian piecewise stationary model of Pham and Cardoso [28]. We show, following the same guidelines as [28], how this likelihood yields a contrast function

for separating dependent sources, which is a joint block diagonalization criterion. We then establish the estimating equations for the mixing matrix \mathbf{A} , that is, the equations satisfied by its maximum likelihood value. Finally, we recast these equations in terms of the component model parameters, the oblique projections \mathbf{P}_i .

A. Piecewise Stationary Model

Let us consider a piecewise stationary model as follows. The observation interval $[1, T]$ is partitioned into Q domains \mathcal{D}_q , $q = 1, \dots, Q$, where domain q contains n_q samples, so that $\sum_{q=1}^Q n_q = T$. We assume that $\mathbf{s}(t)$ is independent of $\mathbf{s}(t')$ if $t \neq t'$ and that, for any $t \in \mathcal{D}_q$, $\mathbf{s}(t)$ has zero mean and covariance matrix $\mathbf{R}_S^{(q)}$. The linear model (1) implies that

$$\mathbf{R}_X^{(q)} = \mathbf{A} \mathbf{R}_S^{(q)} \mathbf{A}^\dagger \quad (8)$$

where $\mathbf{R}_X^{(q)}$ is the covariance matrix of $\mathbf{x}(t)$ for $t \in \mathcal{D}_q$. The empirical counterpart and natural estimate of $\mathbf{R}_X^{(q)}$ is

$$\bar{\mathbf{R}}_X^{(q)} \triangleq \frac{1}{n_q} \sum_{t \in \mathcal{D}_q} \mathbf{x}(t) \mathbf{x}^\dagger(t). \quad (9)$$

The model of dependent sources, discussed in Section II, corresponds to the block-diagonal structure

$$\mathbf{R}_S^{(q)} \triangleq \begin{bmatrix} \mathbf{R}_{S,11}^{(q)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_{S,nn}^{(q)} \end{bmatrix} \quad (10)$$

where $\mathbf{R}_{S,ii}^{(q)}$ is the covariance matrix of $\mathbf{s}_i(t)$ for $t \in \mathcal{D}_q$. The set $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$ is such that it cannot be further jointly block diagonalized into smaller blocks. Hence, $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$ is *JBD-irreducible* (this notion of irreducibility is analogous to that proposed by [16] and [35]).

In the following, we use $\text{bdia}\{\cdot, \dots, \cdot\}$ to denote a block-diagonal matrix constructed from the set of matrices in brackets. Therefore, (10) can be rewritten as

$$\mathbf{R}_S^{(q)} = \text{bdia}\{\mathbf{R}_{S,11}^{(q)}, \dots, \mathbf{R}_{S,nn}^{(q)}\}.$$

We shall also use the related notation $\text{bdia}_{\mathbf{m}}\{\mathbf{M}\}$ which, given an $m \times m$ matrix \mathbf{M} , returns the block-diagonal matrix with block pattern \mathbf{m} which has the same diagonal blocks as \mathbf{M} and has zeros in the off-diagonal blocks.

B. Likelihood

If $\mathbf{s}(t)$ is normally distributed, then the log-likelihood for the model just described is

$$\begin{aligned} \log p(\{\mathbf{x}(t)\}_{t=1}^T) &= \sum_{t=1}^T \log p(\mathbf{x}(t)) \\ &= - \sum_{t=1}^T \frac{1}{2} \left(\log \det 2\pi \mathbf{R}_X^{(t)} + \mathbf{x}^\dagger(t) \mathbf{R}_X^{-1}(t) \mathbf{x}(t) \right) \\ &= - \sum_{q=1}^Q n_q \frac{1}{2} \left(\log \det 2\pi \mathbf{R}_X^{(q)} + \text{tr}\{\bar{\mathbf{R}}_X^{(q)} \mathbf{R}_X^{-1}(q)\} \right) \quad (11) \end{aligned}$$

as we now explain. The first equality comes from the assumption of independence for $t' \neq t$. The second equality follows

from the Gaussian assumption $\mathbf{x}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_X^{(t)})$ and uses the notation $(\mathbf{R}_X^{(t)})^{-1} = \mathbf{R}_X^{-\dagger(t)}$. The third equality results from the piecewise stationary model that $\mathbf{R}^{(t)} = \mathbf{R}^{(q)}$ for $t \in \mathcal{D}_q$ and uses the property $\mathbf{a}^\dagger \mathbf{R} \mathbf{a} = \text{tr}\{\mathbf{R} \mathbf{a} \mathbf{a}^\dagger\}$ for any vector \mathbf{a} and matrix \mathbf{R} of appropriate dimensions.

Using the notation

$$D(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{2} (\text{tr}\{\mathbf{R}_1 \mathbf{R}_2^{-1}\} - \log \det(\mathbf{R}_1 \mathbf{R}_2^{-1}) - m) \quad (12)$$

for any two $m \times m$ positive-definite matrices \mathbf{R}_1 and \mathbf{R}_2 , the log-likelihood (11) can be rewritten as

$$\begin{aligned} \log p(\{\mathbf{x}(t)\}_{t=1}^T) &= - \sum_{q=1}^Q n_q D(\overline{\mathbf{R}}_X^{(q)}, \mathbf{R}_X^{(q)}) + \kappa \\ &= - \sum_{q=1}^Q n_q D(\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}, \mathbf{R}_S^{(q)}) + \kappa \\ &= -T \langle D(\mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger}, \mathbf{R}_S) \rangle + \kappa \quad (13) \end{aligned}$$

where $\kappa = -\frac{1}{2}(mT + \sum_{q=1}^Q n_q \log \det 2\pi \overline{\mathbf{R}}_X^{(q)})$ denotes the term which is irrelevant to the maximization of the likelihood with respect to its parameters, since it depends only on the data, not on the model. The second equality uses (8) and then the invariance of (12) under any invertible transform: for any two positive-definite matrices \mathbf{R}_1 , \mathbf{R}_2 and for any invertible \mathbf{A} ,

$$D(\mathbf{R}_1, \mathbf{A} \mathbf{R}_2 \mathbf{A}^\dagger) = D(\mathbf{A}^{-1} \mathbf{R}_1 \mathbf{A}^{-\dagger}, \mathbf{R}_2). \quad (14)$$

The last step introduces the notation $\langle \cdot \rangle$ to denote a weighted average of any sequence indexed by q with weights n_q :

$$\langle \mathbf{M} \rangle \triangleq \frac{1}{T} \sum_{q=1}^Q n_q \mathbf{M}^{(q)}. \quad (15)$$

C. Contrast Function for a Mixture of Dependent Sources

We can now focus on the case of interest: a mixture of dependent sources. The contrast function is obtained by maximizing the likelihood with respect to the nuisance parameters $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$ for fixed \mathbf{A} . Following the derivation in Appendix B, the ML estimate of $\mathbf{R}_S^{(q)}$ is $\text{bdiag}_{\mathbf{m}}\{\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}\}$ and

$$\max_{\mathbf{R}_S^{(q)}} \log p(\{\mathbf{x}(t)\}_{t=1}^T; \mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) = -T C(\mathbf{A}) + \kappa$$

where we define the *contrast function* [36], [48]

$$C(\mathbf{A}) \triangleq \langle D(\mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger}, \text{bdiag}_{\mathbf{m}}\{\mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger}\}) \rangle \quad (16)$$

and where, for brevity, the dependence of $C(\mathbf{A})$ on the data via $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$ is not denoted explicitly. Note that (16) is the multidimensional analogue of its one-dimensional counterpart in [28].

The scalar $D(\mathbf{R}_1, \mathbf{R}_2)$, defined in (12), is the Kullback–Leibler divergence between the distributions $\mathcal{N}(\mathbf{0}, \mathbf{R}_1)$ and $\mathcal{N}(\mathbf{0}, \mathbf{R}_2)$ [37], and thus is a measure of mismatch between two positive-definite matrices \mathbf{R}_1 and \mathbf{R}_2 . Therefore, in our

piecewise stationary model, maximizing (13) is equivalent to minimizing the average mismatch $\langle D(\overline{\mathbf{R}}_X, \mathbf{R}_X) \rangle$ between the sample covariance matrices and their expected counterparts. Since $D(\mathbf{R}, \text{bdiag}_{\mathbf{m}}\{\mathbf{R}\}) \geq 0$ with equality if and only if \mathbf{R} is block diagonal with block pattern \mathbf{m} , then, for any positive-definite matrix \mathbf{R} , the divergence $D(\mathbf{R}, \text{bdiag}_{\mathbf{m}}\{\mathbf{R}\})$ is a measure of the block-diagonality of \mathbf{R} . Therefore, minimizing $C(\mathbf{A})$ can be understood as joint block diagonalization (JBD) of the set of covariance matrices $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$ by matrix \mathbf{A}^{-1} (see also [38] and [39]).

D. Estimating Equations in Terms of the Mixing Matrix

The next step is to solve for the mixing matrix \mathbf{A} . This is obtained by characterizing the stationary points of the log-likelihood (13) and thus also of the contrast function (16). For this purpose, we calculate the derivative of the likelihood function $\phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) \triangleq \log p(\{\mathbf{x}(t)\}_{t=1}^T)$ with respect to \mathbf{A} , for fixed $\mathbf{R}_S^{(q)}$ (we omit the dependence of ϕ on the data, for brevity). The first-order variation of $\phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q)$ when \mathbf{A} is replaced by $\mathbf{A}(\mathbf{I} + \mathbf{E})$ (where \mathbf{I} denotes the identity matrix) can always be expressed by the Taylor expansion

$$\begin{aligned} \phi(\mathbf{A}(\mathbf{I} + \mathbf{E}), \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) &= \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) \\ &+ \text{tr}\{(\nabla \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q))^\dagger \mathbf{E}\} + \text{higher-order terms in } \mathbf{E}, \end{aligned} \quad (17)$$

for some $m \times m$ matrix $\nabla \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q)$, called the relative gradient of $\phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q)$ with respect to \mathbf{A} . Similarly to the derivation for the one-dimensional case in [28], one obtains that

$$\nabla \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) = \langle \mathbf{R}_S^{-1} \mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger} \rangle - \mathbf{I}. \quad (18)$$

In order to obtain the ML estimate of \mathbf{A} , we equate (18) to zero. Since (18) depends on the nuisance parameters, we can now replace $\mathbf{R}_S^{(q)}$ with its ML estimate, derived in Section III-C. This procedure yields the *estimating equations*

$$\langle \text{bdiag}_{\mathbf{m}}^{-1}\{\mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger}\} \mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger} \rangle = \mathbf{I}. \quad (19)$$

It can be shown that $\nabla C(\mathbf{A})$, the first-order variation of $C(\mathbf{A})$, derived similarly to (17), obeys

$$\nabla \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) \Big|_{\mathbf{R}_S^{(q)} = \text{bdiag}_{\mathbf{m}}\{\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}\}} = -\nabla C(\mathbf{A}).$$

Therefore, the solution of (19) is the ML estimate of \mathbf{A} as well as the stationary points of (16). The estimating equations (19) read block-wise

$$\langle ([\mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger}]_{ii})^{-1} [\mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger}]_{ij} \rangle = \mathbf{0}_{m_i \times m_j} \quad j \neq i \quad (20)$$

where i, j are understood as *block* indices. Using the horizontal blocks \mathbf{B}_i of matrix $\mathbf{B} = \mathbf{A}^{-1}$, (20) is rewritten as

$$\langle (\mathbf{B}_i \overline{\mathbf{R}}_X \mathbf{B}_i^\dagger)^{-1} (\mathbf{B}_i \overline{\mathbf{R}}_X \mathbf{B}_j^\dagger) \rangle = \mathbf{0}_{m_i \times m_j} \quad j \neq i. \quad (21)$$

Note that the (i, i) th block of (19), that is, $i = j$ of (20), degenerates into the identity matrix: the diagonal blocks $i = j$ do not yield any constraints, reflecting the indeterminacy discussed in Section II.

E. Estimating Equations in Terms of the Projectors

The estimating equations (21) can also be expressed as conditions on the oblique projectors \mathbf{P}_i . To do so, we multiply (21) on the left by $\mathbf{A}_i^{\sharp\dagger}$ and on the right by \mathbf{A}_j^\dagger . In the middle, we insert $\mathbf{A}_i^\sharp \mathbf{A}_i = \mathbf{I}$, to obtain

$$\langle \mathbf{A}_i^{\sharp\dagger} (\mathbf{B}_i \overline{\mathbf{R}}_X \mathbf{B}_i^\dagger)^{-1} \mathbf{A}_i^\sharp \mathbf{A}_i (\mathbf{B}_i \overline{\mathbf{R}}_X \mathbf{B}_j^\dagger) \mathbf{A}_j^\dagger \rangle = \mathbf{0}_{m \times m}. \quad (22)$$

We split the expression inside the angular brackets into two factors which can be re-expressed as follows. The leftmost factor is

$$\begin{aligned} \mathbf{A}_i^{\sharp\dagger} (\mathbf{B}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{B}_i^\dagger)^{-1} \mathbf{A}_i^\sharp &= (\mathbf{A}_i \mathbf{B}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{B}_i^\dagger \mathbf{A}_i^\dagger)^\sharp \\ &= (\mathbf{P}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{P}_i^\dagger)^\sharp \end{aligned} \quad (23)$$

where the first equality uses the property that for any invertible $n \times n$ matrix \mathbf{M} and for any $m \times n$ rank- n matrix \mathbf{V} ,

$$\mathbf{V}^{\sharp\dagger} \mathbf{M}^{-1} \mathbf{V}^\sharp = (\mathbf{V} \mathbf{M} \mathbf{V}^\dagger)^\sharp \quad (24)$$

(it is immediate to verify that (24) fulfills all four criteria of the Moore–Penrose pseudoinverse (49)) and the second equality uses definition (6) of the projectors. The rightmost factor is

$$\mathbf{A}_i (\mathbf{B}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{B}_j^\dagger) \mathbf{A}_j^\dagger = \mathbf{P}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{P}_j^\dagger. \quad (25)$$

Substituting (23) and (25) in (22), the estimating equations (21) also read

$$\langle (\mathbf{P}_i \overline{\mathbf{R}}_X \mathbf{P}_i^\dagger)^\sharp (\mathbf{P}_i \overline{\mathbf{R}}_X \mathbf{P}_j^\dagger) \rangle = \mathbf{0}_{m \times m} \quad j \neq i, \quad (26)$$

which is the desired form: the values of the oblique projectors for which the contrast function is stationary are the solutions of the estimating equations (26). Since (20) yields the ML estimates of \mathbf{A} , (26) yields the ML estimates of \mathbf{P}_i . The estimating equations (26) will allow us to lead the error analysis from the components' point of view, as we shall see in Section IV.

IV. ERROR ANALYSIS

We turn to the error analysis of the estimates, obtained by minimizing the contrast function (16). Our purpose is to derive a closed-form expression for the MSE in component estimation. In Section IV-A we define the error in component estimation. We express this error as a function of the error in the oblique projections and of the observations. The error in the oblique projections, which is due to all n components, is decomposed into pairwise error terms. Following a first-order expansion of the estimating equations (26), we obtain in Section IV-B an expression for the pairwise error terms, which depends on the model parameters and on the observations. In Section IV-C, we derive the covariance of the pairwise error terms and thus also of the error in the oblique projections. This derivation provides us with the FIM and CRLB for these estimated parameters, when the Gaussian model holds. Finally, in Section IV-D, based on the former results, we obtain a closed-form expression for the (normalized) MSE for component separation.

We consider an asymptotic analysis in the regime of small errors, in which the results are obtained from a first-order expansion of the estimating equations. In the following, we define asymptotic conditions as $T \rightarrow \infty$ with $\frac{n_q}{T}$ fixed $\forall q$. The

analysis is conducted under the assumption that the model of Section III-A holds.

In order to avoid ambiguities, for certain parameters, a “ \star ” is used to denote the ‘true’ model quantities, and a hat denotes the value which solves the estimating equations. For instance, \mathbf{A}^\star denotes the true mixing matrix, while $\widehat{\mathbf{A}}$ denotes the solution of (20), which is also the ML estimate if the Gaussian model holds.

A. Error Decomposition

A difficulty in error analysis for the multidimensional problem stems from the inability to characterize the estimation error of the mixing matrix, due to the severe indeterminacies it suffers from, as discussed in Section II. We thus begin by defining convenient error terms. In order to focus on well-defined quantities, we consider the errors $\delta \mathbf{P}_i$ in $\widehat{\mathbf{P}}_i$, the ML estimates of the oblique projectors:

$$\delta \mathbf{P}_i = \widehat{\mathbf{P}}_i - \mathbf{P}_i^\star. \quad (27)$$

The estimated i th component is thus

$$\widehat{\mathbf{x}}_i(t) = \widehat{\mathbf{P}}_i \mathbf{x}(t) = (\mathbf{P}_i^\star + \delta \mathbf{P}_i) \mathbf{x}(t) = \mathbf{x}_i(t) + \delta \mathbf{P}_i \mathbf{x}(t)$$

so that, using $\mathbf{x}(t) = \sum_{j=1}^n \mathbf{x}_j(t)$, the error in the i th component is decomposed as

$$\begin{aligned} \widehat{\mathbf{x}}_i(t) - \mathbf{x}_i(t) &= \delta \mathbf{P}_i \mathbf{x}(t) = \delta \mathbf{P}_i \sum_{j=1}^n \mathbf{x}_j(t) \\ &= \sum_{j=1}^n \delta \mathbf{P}_i \mathbf{\Pi}_j^\star \mathbf{x}_j(t) = \sum_{j=1}^n \mathcal{E}_{ij} \mathbf{x}_j(t) \end{aligned} \quad (28)$$

where we have defined $m \times m$ error matrices

$$\mathcal{E}_{ij} \triangleq \delta \mathbf{P}_i \mathbf{\Pi}_j^\star = (\widehat{\mathbf{P}}_i - \mathbf{P}_i^\star) \mathbf{\Pi}_j^\star. \quad (29)$$

The term $\mathbf{\Pi}_j^\star$ in (28) is inserted since it arises naturally in the derivation of the first-order expansion of the estimating equations (Appendix C, (58)); recall that $\mathbf{\Pi}_j^\star \mathbf{x}_j(t) = \mathbf{x}_j(t) \forall j$. The double-indexed term \mathcal{E}_{ij} gives the linearized estimating equations their pairwise form, as will be seen shortly.

For $i \neq j$, the term $\mathcal{E}_{ij} \mathbf{x}_j(t)$ in (28) is called the (i, j) -contamination error, that is, the contamination due to the j th component in the reconstruction of the i th component. The term $\mathcal{E}_{ii} \mathbf{x}_i(t)$ is called the i th-reconstruction error, since this term represents a distortion of $\mathbf{x}_i(t)$ but not any contamination by the other components.

B. Influence Function

In order to evaluate the covariance of the estimation error, we first establish the first-order expansion of \mathcal{E}_{ij} in terms of the finite-sample $m \times m$ covariance matrices

$$\overline{\mathbf{R}}_{X_i X_j}^{(q)} \triangleq \frac{1}{n_q} \sum_{t \in \mathcal{D}_q} \mathbf{x}_i(t) \mathbf{x}_j^\dagger(t). \quad (30)$$

The key assumption for blind separation is block-decorrelation: $\overline{\mathbf{R}}_{X_i X_j}^{(q)} \triangleq E\{\overline{\mathbf{R}}_{X_i X_j}^{(q)}\} = \mathbf{0}_{m \times m}$ for $j \neq i$. However, because of finite sample size, this does not hold for its empirical counterpart, i.e., $\overline{\mathbf{R}}_{X_i X_j}^{(q)} \neq \mathbf{0}_{m \times m}$. In this section, we develop the performance analysis in the regime of small errors, that is, we

analyze the error terms $\widehat{\mathbf{P}}_i - \mathbf{P}_i^*$ at first-order in $\overline{\mathbf{R}}_{X_i X_j}^{(q)}$ when asymptotic conditions hold. From (29), $\boldsymbol{\mathcal{E}}_{ij}$ decreases with T at the same rate as $\delta \mathbf{P}_i$. Assuming that asymptotic conditions hold, then $\widehat{\mathbf{P}}_i \cong \mathbf{P}_i^*$ (see Appendix C).

The first-order expansion of the estimating equations (26) yields (see Appendix C) a set of *pairs* of equations:

$$\begin{aligned} -\langle \mathbf{R}_{X_i X_i}^\# \overline{\mathbf{R}}_{X_i X_j} \rangle &= \boldsymbol{\mathcal{E}}_{ji}^\dagger + \langle \mathbf{R}_{X_i X_i}^\# \boldsymbol{\mathcal{E}}_{ij} \mathbf{R}_{X_j X_j} \rangle + \Omega\left(\frac{1}{T}\right) \\ -\langle \mathbf{R}_{X_j X_j}^\# \overline{\mathbf{R}}_{X_j X_i} \rangle &= \boldsymbol{\mathcal{E}}_{ij}^\dagger + \langle \mathbf{R}_{X_j X_j}^\# \boldsymbol{\mathcal{E}}_{ji} \mathbf{R}_{X_i X_i} \rangle + \Omega\left(\frac{1}{T}\right) \end{aligned} \quad (31)$$

with one such pair of equations for each pair $i \neq j$ of components. Equation (31) shows that asymptotically, for each pair of components, the projector error terms $(\boldsymbol{\mathcal{E}}_{ij}, \boldsymbol{\mathcal{E}}_{ji})$ are related to the corresponding set of $2Q$ matrices $\{\overline{\mathbf{R}}_{X_i X_j}^{(q)}, \overline{\mathbf{R}}_{X_j X_i}^{(q)}\}_{q=1}^Q$, which represents the block-decorrelation error. Such a pairwise decoupling is customary in the asymptotic analysis of ICA algorithms, e.g., [32] and [40].

In order to proceed, it is convenient to vectorize the matrices using the $\text{vec}\{\cdot\}$ operator which stacks the columns of a $p \times q$ matrix into a $pq \times 1$ vector. The pair of (31) can thus be rewritten in matrix form as

$$\begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix} = -\mathcal{H} \cdot \begin{bmatrix} \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} \\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\} \end{bmatrix} + \Omega\left(\frac{1}{T}\right) \quad (32)$$

where

$$\mathbf{g}_{ij} = \langle \text{vec}\{\mathbf{R}_{X_i X_i}^\# \overline{\mathbf{R}}_{X_i X_j}\} \rangle \quad (33)$$

and $\mathcal{H} = \begin{bmatrix} \mathbf{H}_{ij} & \mathbf{T}_{m,m} \\ \mathbf{T}_{m,m} & \mathbf{H}_{ji} \end{bmatrix}$ is a symmetric matrix with

$$\mathbf{H}_{ij} = \langle \mathbf{R}_{X_j X_j} \otimes \mathbf{R}_{X_i X_i}^\# \rangle. \quad (34)$$

In the above, we have introduced the $mn \times mn$ commutation matrix $\mathbf{T}_{m,n}$ [41] such that

$$\text{vec}\{\mathbf{M}^\dagger\} = \mathbf{T}_{m,n} \text{vec}\{\mathbf{M}\} \quad (35)$$

for any $m \times n$ matrix \mathbf{M} . Assuming that \mathcal{H} is invertible (see Section V), then

$$\begin{bmatrix} \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} \\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\} \end{bmatrix} = -\mathcal{H}^{-1} \begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix} + \Omega\left(\frac{1}{T}\right). \quad (36)$$

Equation (36) shows how the empirical correlation between components, that is, the fact that $\overline{\mathbf{R}}_{X_i X_j}^{(q)}$ is non-zero in finite sample size, results in non-zero errors $\boldsymbol{\mathcal{E}}_{ij}$. Note the similarity between (36) and its one-dimensional, source-wise counterpart in [40] and [28]. Equation (36) is the desired closed-form, first-order expression for the error terms in (28).

C. Error Covariance for the Projectors

We are almost ready now to calculate the covariance matrix of $\delta \mathbf{P}_i$, defined in (27). First, we notice that $\delta \mathbf{P}_i = \delta \mathbf{P}_i \sum_{j=1}^n \mathbf{P}_j^* = \sum_{j=1}^n \delta \mathbf{P}_i \Pi_j^* \mathbf{P}_j^* = \sum_{j=1}^n \boldsymbol{\mathcal{E}}_{ij} \mathbf{P}_j^*$, where $\boldsymbol{\mathcal{E}}_{ij}$ are given explicitly by (36) for $i \neq j$. For $\boldsymbol{\mathcal{E}}_{ii}$ one more step is required, because $\boldsymbol{\mathcal{E}}_{ii}$ is not given by (36). Since $\mathbf{x}(t)$ is given, we exploit it by constraining $\sum_{j=1}^n \widehat{\mathbf{x}}_j(t) = \mathbf{x}(t)$. Therefore, $\sum_{j=1}^n \widehat{\mathbf{P}}_j = \mathbf{I}$, which implies $\sum_{j=1}^n \delta \mathbf{P}_j = \mathbf{0}_{m \times m}$, hence $\sum_{j=1}^n \boldsymbol{\mathcal{E}}_{ji} = \mathbf{0}_{m \times m}$. We can thus

rewrite $\delta \mathbf{P}_i = \sum_{j \neq i} (\boldsymbol{\mathcal{E}}_{ij} \mathbf{P}_j^* - \boldsymbol{\mathcal{E}}_{ji} \mathbf{P}_i^*)$. Vectorizing $\delta \mathbf{P}_i$, its covariance matrix is given by

$$\begin{aligned} &\text{Cov}(\text{vec}\{\delta \mathbf{P}_i\}) \\ &= \text{Cov}\left(\sum_{j \neq i} ((\mathbf{P}_j^{*\dagger} \otimes \mathbf{I}) \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} - (\mathbf{P}_i^{*\dagger} \otimes \mathbf{I}) \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\})\right) \\ &= \sum_{j \neq i} ((\mathbf{P}_j^{*\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}) (\mathbf{P}_j^* \otimes \mathbf{I}) \\ &\quad - (\mathbf{P}_j^{*\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}) (\mathbf{P}_i^* \otimes \mathbf{I}) \\ &\quad - (\mathbf{P}_i^{*\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}) (\mathbf{P}_j^* \otimes \mathbf{I}) \\ &\quad + (\mathbf{P}_i^{*\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}) (\mathbf{P}_i^* \otimes \mathbf{I})) + O\left(\frac{1}{T^2}\right) \end{aligned} \quad (37)$$

where the first equality is due to (48c) and the last equality is due to the lack of correlation between the components.

It remains now to calculate the four covariance matrices in (37). These will be taken from the covariance of (36). For this aim, we shall calculate the covariance matrix of the stochastic vector $[\mathbf{g}_{ij}^\dagger, \mathbf{g}_{ji}^\dagger]^\dagger$. We show in Appendix D that

$$\text{Cov}\left(\begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix}\right) = \frac{1}{T} \mathcal{H}_\Pi \quad (38)$$

where

$$\mathcal{H}_\Pi = \begin{bmatrix} \mathbf{H}_{ij} & \mathbf{T}_{m,m}^{\Pi_{ij}} \\ \mathbf{T}_{m,m}^{\Pi_{ji}} & \mathbf{H}_{ji} \end{bmatrix} \quad (39)$$

and

$$\mathbf{T}_{m,m}^{\Pi_{ij}} \triangleq (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*) \mathbf{T}_{m,m} (\mathbf{\Pi}_i^* \otimes \mathbf{\Pi}_j^*). \quad (40)$$

Using (38), the covariance matrix of (36), for any $i \neq j$, is given by

$$\begin{aligned} \text{Cov}\left(\begin{bmatrix} \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} \\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\} \end{bmatrix}\right) &= \frac{1}{T} \mathcal{H}^{-1} \mathcal{H}_\Pi \mathcal{H}^{-\dagger} + O\left(\frac{1}{T^2}\right) \\ &= \frac{1}{T} \mathcal{H}^\# + O\left(\frac{1}{T^2}\right) \end{aligned} \quad (41)$$

where the last step is proved in Appendix E. Further manipulations on (41) (Appendix F) yield, for $i \neq j$,

$$\text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}) \quad (42)$$

$$= \frac{1}{T} (\langle \mathbf{R}_{X_j X_j} \otimes \mathbf{R}_{X_i X_i}^\# \rangle - \langle \mathbf{R}_{X_j X_j}^\# \otimes \mathbf{R}_{X_i X_i} \rangle)^\# + O\left(\frac{1}{T^2}\right)$$

and

$$\text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}) \quad (43)$$

$$= \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}) \cdot \langle \mathbf{R}_{X_j X_j}^\# \otimes \mathbf{R}_{X_i X_i} \rangle^\# \mathbf{T}_{m,m} + O\left(\frac{1}{T^2}\right)$$

as the cross-covariance matrix between $\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}$ and $\text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}$. All the covariance matrices on the RHS of (37) are now given in explicit form by (42) and (43), which concludes the closed-form derivation of $\text{Cov}(\text{vec}\{\delta \mathbf{P}_i\})$.

Under the Gaussian assumption, the results in this section have the following interpretation. Since \mathbf{g}_{ij} is the first-order expansion of the relative gradient (26) of the log-likelihood (11), then (32) with (38) imply that \mathcal{H}_Π is the FIM for the pair $(\text{vec}\{\mathbf{P}_i^* \mathbf{\Pi}_j^*\}, \text{vec}\{\mathbf{P}_j^* \mathbf{\Pi}_i^*\})$, whose estimation error is given by $(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\})$ (see (29)). In (41), we obtain that the covariance matrix of the estimation errors is (approximately) equal to the pseudoinverse of the FIM. Therefore, (41) is the

asymptotically achievable CRLB on the estimation of the pair $(\text{vec}\{\mathbf{P}_i^* \mathbf{\Pi}_j^*\}, \text{vec}\{\mathbf{P}_j^* \mathbf{\Pi}_i^*\})$, and (42) is the CRLB for $\text{vec}\{\mathbf{P}_i^* \mathbf{\Pi}_j^*\}$ alone. Since we have shown that (42) is the CRLB for $\delta \mathbf{P}_i \mathbf{\Pi}_j^*$, and \mathbf{P}_i is a linear function thereof (up to $O(\frac{1}{T^2})$ terms), (37) is the CRLB for $\widehat{\mathbf{P}}_i$.

D. Mean Square Error

The final step in our analysis is to propagate expression (37) for the covariance matrix of the oblique projection matrices $\widehat{\mathbf{P}}_i$ into an expression for the component estimation error. Let us define the estimation error of a given component i by a normalized MSE:

$$\widehat{\text{MSE}}_i \triangleq \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^T |\widehat{\mathbf{x}}_i(t) - \mathbf{x}_i(t)|^2 \quad (44)$$

where the normalization is by the average power σ_i^2 of the i th component,

$$\sigma_i^2 \triangleq \frac{1}{T} \sum_{t=1}^T E \{ |\mathbf{x}_i(t)|^2 \} = \text{tr} \{ \langle \mathbf{R}_{X_i X_i} \rangle \}, \quad (45)$$

and where the last step uses $|\mathbf{x}_i(t)|^2 = \text{tr} \{ \mathbf{x}_i(t) \mathbf{x}_i^\dagger(t) \}$ and (9). Using (28), (44) can now be rewritten as

$$\begin{aligned} \widehat{\text{MSE}}_i &= \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^T |\delta \mathbf{P}_i \mathbf{x}(t)|^2 \\ &= \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^T \text{tr} \{ \delta \mathbf{P}_i \mathbf{x}(t) \mathbf{x}^\dagger(t) (\delta \mathbf{P}_i)^\dagger \} \\ &= \frac{1}{\sigma_i^2} \text{tr} \{ (\langle \mathbf{R}_X \rangle \otimes \mathbf{I}) \text{vec} \{ \delta \mathbf{P}_i \} \text{vec}^\dagger \{ \delta \mathbf{P}_i \} \} \quad (46) \end{aligned}$$

where the second equality is analogous to that in (45), and the last equality uses (9) and then Property A.1 in Appendix A.

Now, in order to obtain the expectation of (46), we employ the reasonable assumption, that $\widehat{\mathbf{P}}_i$, and thus $\delta \mathbf{P}_i$, are statistically independent of the *total* power of the observations, $\langle \mathbf{R}_X \rangle$. This follows from the fact that reliable JBD cannot be obtained with less than two matrices, see Section V. Hence, we can write

$$\begin{aligned} \text{MSE}_i &\triangleq E \{ \widehat{\text{MSE}}_i \} \\ &= \frac{1}{\sigma_i^2} \text{tr} \{ (\langle \mathbf{R}_X \rangle \otimes \mathbf{I}) \text{Cov}(\text{vec} \{ \delta \mathbf{P}_i \}) \} \quad (47) \end{aligned}$$

where $\text{Cov}(\text{vec} \{ \delta \mathbf{P}_i \})$ is given in (37), with (42) and (43). Note that due to (48d) and the statistical independence between any two different components, only the first and fourth summands on the RHS of (37) contribute to (47). We have thus obtained a closed-form expression for the MSE, which can be fully expressed by the model parameters $\{ \mathbf{R}_{X_i X_i}^{(q)} \}_{q=1}^Q$ and the weights $\{ n_q \}_{q=1}^Q$, up to $O(\frac{1}{T^2})$ terms. In the Gaussian case, this is also the minimal MSE (MMSE).

It should be noted that all the derivations in this Section IV and in the related appendices do not rely neither on Gaussian distribution nor on statistics of order larger than 2. Therefore, (42), (43), (37) and (47) hold also for non-Gaussian observations. That is, they still reflect the error covariance and MSE if we apply (16) for their separation; however, the CRLB, FIM

and MMSE interpretation of the derived expressions no longer applies.

V. IDENTIFIABILITY

In this section, we 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. Since we focus on second-order methods, the data are represented only by their sample covariance matrices. These are Q symmetric $m \times m$ matrices so that our model should try to fit

$$N_{\text{data}} = Q \frac{m(m+1)}{2}$$

scalar numbers. The model is adjusted by varying the mixing matrix and the source covariance matrices. However, there is some redundancy between these matrices, because of the factorizations discussed in Section II: each submatrix \mathbf{A}_i has $m \times m_i$ degrees of freedom, where m_i^2 of them can be factored into the corresponding source covariance matrices. This leaves $m^2 - \sum_{i=1}^n m_i^2 \geq 0$ effective degrees of freedom in the mixing matrix \mathbf{A} , and $m_i(m_i + 1)/2$ degrees of freedom in each $\mathbf{R}_{S,ii}^{(q)}$. Hence, the model has

$$N_{\text{model}} = m^2 - \sum_{i=1}^n m_i^2 + Q \sum_{i=1}^n \frac{m_i(m_i + 1)}{2}$$

effective free scalar parameters. It turns out that

$$N_{\text{data}} - N_{\text{model}} = \left(\frac{Q}{2} - 1 \right) \left(m^2 - \sum_{i=1}^n m_i^2 \right).$$

Hence, as soon as $Q \geq 2$, we have $N_{\text{data}} \geq N_{\text{model}}$, that is, there are more (or, at least, as many) scalar statistics as free parameters in the model.

However, since *any* two positive-definite matrices can be *exactly* jointly diagonalized [42, Theorem 6], the JBD-irreducibility requirement of Section III-A will be violated if we let $Q = 2$ for multidimensional data. The latter assumption implies that $Q > 2$ is a necessary condition for identifiability in the presence of multidimensional components. Otherwise, one can suffice with $Q = 2$.

B. Identifiability and Uniqueness

The previous argument makes it plausible that for randomly chosen source covariance matrices, the component subspaces can be identified blindly. As shown in [43], existence and uniqueness of JBD, up to trivial indeterminacies (see Section II) and for $Q > 2$ is guaranteed with probability one when the entries of the source covariance matrices are drawn from a continuous probability density function. It thus remains to specify when this uniqueness does not hold. For one-dimensional ICA, it has been shown by [44] that non-identifiability occurs when the covariances of the different sources, as a function of the domain index, are proportional. Recently, [45] have developed analogous conditions for the multidimensional case, in terms of the source covariance matrices $\mathbf{R}_S^{(q)}$. We thus assume that all the required conditions for uniqueness and

TABLE I
 PERFORMANCE OF SECOND-ORDER MULTIDIMENSIONAL ICA: ANALYTICAL VERSUS EMPIRICAL NORMALIZED MSE, AVERAGED OVER 5000 MONTE-CARLO TRIALS. EMPIRICAL RESULTS: JBD ALGORITHM WITH CORRECT \mathbf{m} ($\widehat{\text{MSE}}_i$) VERSUS $\mathbf{m} = [1, \dots, 1]^\dagger$ ($\widehat{\text{MSE}}_i^{\text{JD}}$). EACH SCENARIO, THAT IS, DIFFERENT \mathbf{m} , IS TESTED ONCE WITH GAUSSIAN AND ONCE WITH NON-GAUSSIAN DATA. THE LAST ROW OF THE TABLE SUMMARIZES THE COLUMNS

Scenario	i	m_i	MSE_i	$\widehat{\text{MSE}}_i$	$\frac{\widehat{\text{MSE}}_i}{\text{MSE}_i}$	$\widehat{\text{MSE}}_i^{\text{JD}}$	$\frac{\widehat{\text{MSE}}_i^{\text{JD}}}{\widehat{\text{MSE}}_i}$	Non-Gaussian Data Type	$\widehat{\text{MSE}}_i$	$\frac{\widehat{\text{MSE}}_i}{\text{MSE}_i}$	$\widehat{\text{MSE}}_i^{\text{JD}}$	$\frac{\widehat{\text{MSE}}_i^{\text{JD}}}{\widehat{\text{MSE}}_i}$
Model			Gaussian data					Non-Gaussian data				
#1	1	2	$1.40 \cdot 10^{-3}$	$1.40 \cdot 10^{-3}$	0.999	$2.11 \cdot 10^{-3}$	1.507	U	$1.40 \cdot 10^{-3}$	0.999	$2.08 \cdot 10^{-3}$	1.484
	2	2	$8.31 \cdot 10^{-4}$	$8.40 \cdot 10^{-4}$	1.010	$1.18 \cdot 10^{-3}$	1.410		$8.22 \cdot 10^{-4}$	0.989	$1.16 \cdot 10^{-3}$	1.415
	3	2	$1.57 \cdot 10^{-3}$	$1.56 \cdot 10^{-3}$	0.993	$2.46 \cdot 10^{-3}$	1.578		$1.58 \cdot 10^{-3}$	1.007	$2.43 \cdot 10^{-3}$	1.539
#2	1	6	$1.58 \cdot 10^{-4}$	$1.56 \cdot 10^{-4}$	0.991	$4.84 \cdot 10^{-4}$	3.099	L	$1.60 \cdot 10^{-4}$	1.018	$5.01 \cdot 10^{-4}$	3.125
	2	1	$4.52 \cdot 10^{-3}$	$4.60 \cdot 10^{-3}$	1.018	$1.70 \cdot 10^{-2}$	3.687		$4.62 \cdot 10^{-3}$	1.022	$1.78 \cdot 10^{-2}$	3.852
	3	1	$2.54 \cdot 10^{-3}$	$2.50 \cdot 10^{-3}$	0.985	$5.91 \cdot 10^{-3}$	2.364		$2.58 \cdot 10^{-3}$	1.019	$6.06 \cdot 10^{-3}$	2.343
#3	1	3	$8.22 \cdot 10^{-4}$	$8.31 \cdot 10^{-4}$	1.010	$1.44 \cdot 10^{-3}$	1.737	U	$8.20 \cdot 10^{-4}$	0.998	$1.42 \cdot 10^{-3}$	1.733
	2	2	$8.87 \cdot 10^{-4}$	$9.07 \cdot 10^{-4}$	1.022	$1.36 \cdot 10^{-3}$	1.500		$8.83 \cdot 10^{-4}$	0.996	$1.35 \cdot 10^{-3}$	1.524
	3	1	$2.60 \cdot 10^{-3}$	$2.59 \cdot 10^{-3}$	0.996	$5.09 \cdot 10^{-3}$	1.961		$2.63 \cdot 10^{-3}$	1.009	$4.98 \cdot 10^{-3}$	1.895
#4	1	5	$4.40 \cdot 10^{-4}$	$4.44 \cdot 10^{-4}$	1.009	$2.49 \cdot 10^{-3}$	5.595	L	$4.40 \cdot 10^{-4}$	0.998	$2.46 \cdot 10^{-3}$	5.585
	2	4	$5.85 \cdot 10^{-4}$	$5.90 \cdot 10^{-4}$	1.008	$3.30 \cdot 10^{-3}$	5.598		$5.85 \cdot 10^{-4}$	0.999	$3.26 \cdot 10^{-3}$	5.581
#5	1	1	$9.67 \cdot 10^{-4}$	$9.72 \cdot 10^{-4}$	1.005	$9.72 \cdot 10^{-4}$	1	GM	$9.58 \cdot 10^{-4}$	0.991	$9.58 \cdot 10^{-4}$	1
	2	1	$2.30 \cdot 10^{-3}$	$2.31 \cdot 10^{-3}$	1.004	$2.31 \cdot 10^{-3}$	1		$2.27 \cdot 10^{-3}$	0.990	$2.27 \cdot 10^{-3}$	1
	3	1	$1.25 \cdot 10^{-4}$	$1.25 \cdot 10^{-4}$	0.998	$1.25 \cdot 10^{-4}$	1		$1.26 \cdot 10^{-4}$	1.007	$1.26 \cdot 10^{-4}$	1
			$\ll 1$	$\ll 1$	$\cong 1$	$\ll 1$	≥ 1		$\ll 1$	$\cong 1$	$\ll 1$	≥ 1

identifiability, as given by [43] and [45], hold. We postulate that these conditions suffice for \mathcal{H} to be invertible, as the derivation of the singular points of \mathcal{H} is beyond the scope of this paper.

VI. NUMERICAL RESULTS

In this section, we validate experimentally the performance analysis of Section IV. Two algorithms which minimize (16), and thus solve (32), are suggested in [30]. Both converge to the same separation point. For the following simulations we preferred the quasi-Newton realization, due to its faster convergence rate.

In the following simulation, we construct the data so that the analysis requirements of Section IV hold, including small-error regime. Therefore, the theoretical prediction of the MSE is expected to be an accurate prediction of the measured error. We set $Q = 5$ adjacent domains with $n_q = 500$ samples for a total of $T = 2500$ samples. In each experiment, the matrices $\mathbf{R}_S^{(q)}$ are drawn as $\mathbf{R}_{S,ii}^{(q)} = \mathbf{U}^\dagger \mathbf{U}$, where \mathbf{U} is an upper triangular matrix with independent entries uniformly distributed on $[-\frac{1}{2}, \frac{1}{2}]$. The underlying sources are created by left-multiplying the Cholesky factorization of $\mathbf{R}_S^{(q)}$ with $m \times n_q$ statistically independent, zero mean, unit variance numbers. These numbers are drawn from various distributions, in order to validate our claim that the second-order analysis holds not only for Gaussian sources.

As explained in Section II, there is no scale indeterminacy to resolve. Since the JBD algorithms [30] do not guarantee global convergence, the following steps were taken in order to avoid permutation errors. We chose to initialize the JBD algorithm [30] with \mathbf{I} . In this case, permutation errors are avoided by choosing mixing matrices \mathbf{A} which are strictly diagonally-dominant. In the following simulations, $\mathbf{A}_{ij} = 3\delta_{ij} + \frac{1}{10}r_{ij}$, $r_{ij} \sim$

$\mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$ and r_{ij} are i.i.d. Such values allow for sufficient variability of the mixing matrix to test our small-error analysis, while maintaining global convergence. Cases in which an algorithm does converge to an undesired local minimum are due to permutation errors. These are easily detected, since they result in a significantly larger MSE. Therefore, as a final safety measure, we verified that no such large errors appeared in our results.

Table I compares the empirical with the analytical MSE for several scenarios with varying component dimensions and distributions. The second column states the arbitrary index given to each component. The third column denotes the dimension of the i th component in the scenario. In each scenario, different \mathbf{A} and $\mathbf{R}_S^{(q)}$ are drawn. The fourth column gives the analytical MSE for each component (47), which is calculated using the correct model parameters. Each scenario is evaluated using 5000 Monte Carlo trials. For each scenario, two data types were tested. In columns 5–8, Gaussian, zero mean, unit-variance numbers are used to create the underlying sources. In columns 10–13, either uniform, Laplacian, or Gaussian mixture (peaks centered at ± 4) zero-mean, unit-variance numbers, denoted U, L, and GM, respectively, are used to create the underlying sources. Note that left-multiplication of the non-Gaussian numbers with the Cholesky factorization of $\mathbf{R}_S^{(q)}$ changes their distribution; however, it is still non-Gaussian. The non-Gaussian distribution, used to generate the data for each scenario, is given in column 9.

The fifth and tenth columns give the averaged empirical MSE for each component (44). Columns 6 and 11 give the ratio of MSE for component separation: simulated versus analytical. Columns 7–8 and 12–13 compare the averaged empirical MSE using our JBD criterion with the averaged empirical MSE obtained from one-dimensional modeling (JD) and then grouping the separated elements into the multidimensional components,

according to the known partition \mathbf{m} . These values are denoted in Table I as $\widehat{\text{MSE}}^{\text{JD}}$. We point out that the conditions derived in [20] for the global optimum to be achieved by properly grouping the ICA elements refer to a *different* separation criterion and are thus inapplicable here.

The last row of Table I summarizes the results of each column. First, note that all the (normalized) MSE values are much smaller than 1, illustrating the quality of the component separation. Second, note that all values in column 6 and 11 are close to 1, showing that our analysis predicts correctly the achievable separation accuracy. The good match between predicted and empirical MSE demonstrates that indeed only second-order statistics are required for our theoretical analysis, and that for Gaussian data, the CRLB is indeed achievable. As expected [34], there is a significant gain (columns 8 and 13) due to using the correct model, as proposed in this paper. An important result is that in scenarios 2 and 3, which include one-dimensional components along with higher-dimensional ones, the gain for the one-dimensional components is >1 , as well. Obviously, when the sources are independent, as in scenario 5, there is no difference between JD and JBD, hence we put “1” in this cell.

VII. CONCLUSION

In this paper, we presented the concept of BSS of multi-dimensional components as a new perspective on the dependent sources model. Based on a piecewise stationary model, we derived an ML-based criterion (16) which singles out the multidimensional, unambiguous components from their sum. This criterion can be interpreted as the JBD of a set of covariance matrices. Error analysis of this criterion provided us with a closed-form expression of the covariance of the oblique projections, which are the unambiguous counterparts of the mixing matrix in the component representation. Our error analysis reveals that for Gaussian data, our separation criterion achieves, up to higher-order terms, the CRLB, and is thus optimal in the MSE sense (MMSE). We then derived a closed-form expression for the MSE of the component estimates, in terms of the covariance matrices of the components. This expression is valid, though no longer optimal, also for non-Gaussian data, when the other model assumptions hold. Our derivations were verified in numerical simulations. The performance of JBD was compared with that of classical one-dimensional joint diagonalization and our treatment of multidimensional components was shown to yield a significant gain in their separation.

APPENDIX A SOME ALGEBRAIC PROPERTIES

For ease of reference, we list some useful algebraic properties. Properties which are not proved below can be found in [41], [46], and [47].

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

$$(\mathbf{N} \otimes \mathbf{M})(\mathbf{P} \otimes \mathbf{Q}) = \mathbf{NP} \otimes \mathbf{MQ} \quad (48a)$$

$$(\mathbf{N} \otimes \mathbf{M})^\dagger = \mathbf{N}^\dagger \otimes \mathbf{M}^\dagger \quad (48b)$$

$$\text{vec}\{\mathbf{MQN}\} = (\mathbf{N}^\dagger \otimes \mathbf{M})\text{vec}\{\mathbf{Q}\} \quad (48c)$$

$$\text{tr}\{\mathbf{PQ}\} = \text{tr}\{\mathbf{QP}\} \quad (48d)$$

$$\text{tr}\{\mathbf{P}^\dagger \mathbf{Q}\} = \text{vec}^\dagger\{\mathbf{P}\}\text{vec}\{\mathbf{Q}\}. \quad (48e)$$

Definition A.1: The Moore–Penrose pseudoinverse of a matrix \mathbf{M} is the matrix $\mathbf{M}^\#$ that obeys

$$\mathbf{MM}^\# \mathbf{M} = \mathbf{M} \quad (49a)$$

$$\mathbf{M}^\# \mathbf{MM}^\# = \mathbf{M}^\# \quad (49b)$$

$$\mathbf{MM}^\# = (\mathbf{MM}^\#)^\dagger \quad (49c)$$

$$\mathbf{M}^\# \mathbf{M} = (\mathbf{M}^\# \mathbf{M})^\dagger. \quad (49d)$$

Property A.1: Let $\mathbf{V}, \mathbf{M}, \mathbf{N}$ be $n \times n$ matrices, \mathbf{V} symmetric. Then

$$\text{tr}\{\mathbf{MVN}^\dagger\} = \text{tr}\{(\mathbf{V} \otimes \mathbf{I})\text{vec}\{\mathbf{M}\}\text{vec}^\dagger\{\mathbf{N}\}\}.$$

Proof: Starting from the LHS,

$$\begin{aligned} \text{tr}\{\mathbf{MVN}^\dagger\} &= \text{vec}^\dagger\{\mathbf{N}\}\text{vec}\{\mathbf{MV}\} \\ &= \text{vec}^\dagger\{\mathbf{N}\}(\mathbf{V}^\dagger \otimes \mathbf{I})\text{vec}\{\mathbf{M}\} \\ &= \text{tr}\{\text{vec}^\dagger\{\mathbf{N}\}(\mathbf{V}^\dagger \otimes \mathbf{I})\text{vec}\{\mathbf{M}\}\} \end{aligned}$$

where the first equality uses (48d) and then (48e). In the second equality, $\text{vec}\{\mathbf{MV}\} = \text{vec}\{\mathbf{IMV}\}$, followed by (48c). In the third equality, we simply add the trace notation. The desired result follows from (48d). \square

APPENDIX B DERIVATION OF THE CONTRAST FUNCTION

This appendix details the steps to obtain the contrast function (16), by maximizing the log-likelihood (13) with respect to $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$. This maximization can be obtained using Property B.1 below, which implies that

$$\min_{\mathbf{N} \in \text{bdiag}_m} D(\mathbf{M}, \mathbf{N}) = D(\mathbf{M}, \text{bdiag}_m\{\mathbf{M}\}). \quad (50)$$

Therefore, for any domain q ,

$$\begin{aligned} \min_{\mathbf{R}_S^{(q)}} D(\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}, \mathbf{R}_S^{(q)}) \\ = D(\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}, \text{bdiag}_m\{\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}\}) \end{aligned}$$

due to (50), since $\mathbf{R}_S^{(q)}$ is block-diagonal by definition. Since the samples are statistically independent, the log-likelihood (13) can be maximized by optimizing each domain separately, which concludes the derivation.

Property B.1 (Closest Block Diagonal Matrix): Given pattern \mathbf{m} , the best, in the Kullback–Leibler divergence sense, block-diagonal approximation to a positive-definite matrix \mathbf{M} , is $\text{bdiag}_m\{\mathbf{M}\}$:

$$\arg \min_{\mathbf{N} \in \text{bdiag}_m} D(\mathbf{M}, \mathbf{N}) = \text{bdiag}_m\{\mathbf{M}\}$$

from which (50) follows.

Proof: For a given pattern \mathbf{m} , let \mathbf{M} and $\mathbf{N} \triangleq \text{bdiag}_m\{\mathbf{N}\}$ be any two positive-definite matrices. Then, the Kullback–Leibler divergence between \mathbf{M} and $\text{bdiag}_m\{\mathbf{N}\}$ can be decomposed into

$$\begin{aligned} D(\mathbf{M}, \text{bdiag}_m\{\mathbf{N}\}) &= D(\mathbf{M}, \text{bdiag}_m\{\mathbf{M}\}) \\ &\quad + D(\text{bdiag}_m\{\mathbf{M}\}, \text{bdiag}_m\{\mathbf{N}\}). \quad (51) \end{aligned}$$

The proof of (51) is straightforward using the definition (12) of the Kullback–Leibler divergence. In (51), $D(\mathbf{M}, \text{bdiag}_m\{\mathbf{M}\})$ is independent of \mathbf{N} . Setting $\mathbf{N} = \text{bdiag}_m\{\mathbf{M}\}$ zeroes (and minimizes, since $D(\cdot, \cdot) \geq 0$) the rightmost term of (51). \square

APPENDIX C

FIRST-ORDER EXPANSION OF THE ESTIMATING EQUATIONS

In this appendix, we show that the first-order expansion of the estimating equations (26) leads to the linear relation (31) between the projection error terms $\mathcal{E}_{ij} = \delta\mathbf{P}_i\Pi_j^\dagger$ and the sample error terms $\overline{\mathbf{R}}_{X_iX_j}^{(q)}$.

Let us begin by restating that the estimates $\widehat{\mathbf{P}}_i$ are solutions of the estimating equations (26). That is, in the notation of Section IV,

$$\langle\langle \widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X \widehat{\mathbf{P}}_i^\dagger \rangle\rangle^\# (\widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X \widehat{\mathbf{P}}_j^\dagger) = \mathbf{0}_{m \times m}. \quad (52)$$

In the following, we linearize these equations with respect to the error terms due to finite sample size,

$$\delta\mathbf{R}_X^{(q)} = \overline{\mathbf{R}}_X^{(q)} - \mathbf{R}_X^{(q)}. \quad (53)$$

Under asymptotic conditions, which are defined formally in Section IV, $\overline{\mathbf{R}}_X^{(q)}$ converges, in the mean square, to $\mathbf{R}_X^{(q)}$, and the ML estimator $\widehat{\mathbf{P}}_i$ converges, in probability, to \mathbf{P}_i^* (asymptotically, for non-Gaussian components, both converge in probability). As for the rate of convergence, the entries of both $\delta\mathbf{R}_X^{(q)}$ and $\delta\mathbf{P}_i$ are zero mean random variables with a standard deviation proportional to $1/\sqrt{T}$. Hence, asymptotically, terms which are proportional to $\delta\mathbf{R}_X^{(q)}$ or $\delta\mathbf{P}_i$ are $\Omega(\frac{1}{\sqrt{T}})$ (this notation is defined in Section I) and are considered as having the same order of magnitude.

Expanding the right-hand term within the $\langle \cdot \rangle$ on the LHS of (52),

$$\begin{aligned} & \widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_j^\dagger \\ &= (\mathbf{P}_i^* + \delta\mathbf{P}_i)(\mathbf{R}_X^{(q)} + \delta\mathbf{R}_X^{(q)})(\mathbf{P}_j^{*\dagger} + \delta\mathbf{P}_j^\dagger) \\ &= \overline{\mathbf{R}}_{X_iX_j}^{(q)} + \mathbf{P}_i^* \mathbf{R}_X^{(q)} \delta\mathbf{P}_j^\dagger + \delta\mathbf{P}_i \mathbf{R}_X^{(q)} \mathbf{P}_j^{*\dagger} + \Omega(\frac{1}{T}) \\ &= \overline{\mathbf{R}}_{X_iX_j}^{(q)} + \mathbf{R}_{X_iX_i}^{(q)} \delta\mathbf{P}_j^\dagger + \delta\mathbf{P}_i \mathbf{R}_{X_jX_j}^{(q)} + \Omega(\frac{1}{T}) \end{aligned} \quad (54)$$

where in the last transition we used $\mathbf{P}_i^* \mathbf{R}_X^{(q)} = \mathbf{R}_{X_iX_i}^{(q)}$. Note that all the terms on the RHS of (54) are $\Omega(\frac{1}{\sqrt{T}})$, that is, first-order terms, since $\mathbf{P}_i^* \mathbf{R}_X^{(q)} \mathbf{P}_j^{*\dagger} = \mathbf{R}_{X_iX_j}^{(q)} = \mathbf{0}_{m \times m}$.

As for the left-hand term within the $\langle \cdot \rangle$ on the LHS of (52),

$$\begin{aligned} & \langle \widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_i^\dagger \rangle^\# = \langle \widehat{\mathbf{A}}_i \widehat{\mathbf{B}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{B}}_i^\dagger \widehat{\mathbf{A}}_i^\dagger \rangle^\# \\ &= \widehat{\mathbf{A}}_i^\dagger \langle \widehat{\mathbf{B}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{B}}_i^\dagger \rangle^{-1} \widehat{\mathbf{A}}_i^\# \\ &= \mathbf{A}_i^\dagger \langle \mathbf{B}_i \mathbf{R}_X^{(q)} \mathbf{B}_i^\dagger \rangle^{-1} \mathbf{A}_i^\# + \Omega(\frac{1}{\sqrt{T}}) \\ &= (\mathbf{P}_i^* \mathbf{R}_X^{(q)} \mathbf{P}_i^{*\dagger})^\# + \Omega(\frac{1}{\sqrt{T}}). \end{aligned} \quad (55)$$

The first equality in (55) is due to $\widehat{\mathbf{P}}_i \triangleq \widehat{\mathbf{A}}_i \widehat{\mathbf{B}}_i$, which follows from (6). The second equality follows from (24). The third transition is due to the fact that $\overline{\mathbf{R}}_X^{(q)}$ and the ML estimates of \mathbf{A}

and \mathbf{B} converge to their mean with a standard deviation proportional to $1/\sqrt{T}$. The last step follows again from (24) and then (6). Multiplying (55) with (54),

$$\begin{aligned} & \langle \widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_i^\dagger \rangle^\# \langle \widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_j^\dagger \rangle \\ &= \mathbf{R}_{X_iX_i}^\# (\overline{\mathbf{R}}_{X_iX_j}^{(q)} + \mathbf{R}_{X_iX_i}^{(q)} \delta\mathbf{P}_j^\dagger + \delta\mathbf{P}_i \mathbf{R}_{X_jX_j}^{(q)}) + \Omega(\frac{1}{T}) \\ &= \mathbf{R}_{X_iX_i}^\# \overline{\mathbf{R}}_{X_iX_j}^{(q)} + \Pi_i^* \delta\mathbf{P}_j^\dagger + \mathbf{R}_{X_iX_i}^\# \delta\mathbf{P}_i \mathbf{R}_{X_jX_j}^{(q)} + \Omega(\frac{1}{T}) \end{aligned} \quad (56)$$

where in the first transition we used $\mathbf{R}_{X_iX_i}^\# = (\mathbf{P}_i^* \mathbf{R}_X^{(q)} \mathbf{P}_i^{*\dagger})^\#$ and in the second one,

$$\mathbf{R}_{X_iX_i}^\# \mathbf{R}_{X_iX_i}^{(q)} = \Pi_i^*. \quad (57)$$

Averaging (56) over all domains, (52) is linearized as

$$\langle \mathbf{R}_{X_iX_i}^\# \overline{\mathbf{R}}_{X_iX_j} \rangle + \Pi_i^* \delta\mathbf{P}_j^\dagger + \langle \mathbf{R}_{X_iX_i}^\# \delta\mathbf{P}_i \mathbf{R}_{X_jX_j} \rangle = \mathbf{0}_{m \times m} + \Omega(\frac{1}{T}). \quad (58)$$

Using the property $\mathbf{R}_{X_iX_i}^{(q)} = \Pi_i^* \mathbf{R}_{X_iX_i}^{(q)}$ and the notation \mathcal{E}_{ij} of (29), (58) is rewritten as

$$\langle \mathbf{R}_{X_iX_i}^\# \overline{\mathbf{R}}_{X_iX_j} \rangle + \mathcal{E}_{ji}^\dagger + \langle \mathbf{R}_{X_iX_i}^\# \mathcal{E}_{ij} \mathbf{R}_{X_jX_j} \rangle = \mathbf{0}_{m \times m} + \Omega(\frac{1}{T}). \quad (59)$$

We have thus proved the first equation in the pair (31). The second equation is obtained by interchanging i and j .

APPENDIX D

CLOSED-FORM EXPRESSION FOR (38)

In this appendix we derive the expression (38) for the covariance matrix of the gradients \mathbf{g}_{ij} defined by (33). Since, by the assumptions in Section III, these gradients have zero mean,

$$\text{Cov} \left(\begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix} \right) = \begin{bmatrix} E\{\mathbf{g}_{ij}\mathbf{g}_{ij}^\dagger\} & E\{\mathbf{g}_{ij}\mathbf{g}_{ji}^\dagger\} \\ E\{\mathbf{g}_{ji}\mathbf{g}_{ij}^\dagger\} & E\{\mathbf{g}_{ji}\mathbf{g}_{ji}^\dagger\} \end{bmatrix}.$$

We shall now prove that

$$E\{\mathbf{g}_{ij}\mathbf{g}_{ij}^\dagger\} = \frac{1}{T} \mathbf{H}_{ij} \quad (60)$$

$$E\{\mathbf{g}_{ij}\mathbf{g}_{ji}^\dagger\} = \frac{1}{T} \mathbf{T}_{m,m}^{\Pi_{ij}} \quad (61)$$

for any $i \neq j$, which provide the desired result.

In a first step, we relate the covariance of the gradients to the covariance of the empirical matrices. For any i, j, k, l ,

$$\begin{aligned} E\{\mathbf{g}_{ij}\mathbf{g}_{kl}^\dagger\} &= \frac{1}{T} \sum_{q=1}^Q n_q \frac{1}{T} \sum_{p=1}^Q n_p (\mathbf{I} \otimes \mathbf{R}_{X_iX_i}^{(q)\#}) \\ &\quad \cdot E\{\text{vec}\{\overline{\mathbf{R}}_{X_iX_j}^{(q)}\} \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_kX_l}^{(p)}\}\} (\mathbf{I} \otimes \mathbf{R}_{X_kX_k}^{(p)\#}) \end{aligned}$$

where we have used the alternate form of (33),

$$\mathbf{g}_{ij} = \langle \text{vec}\{\mathbf{R}_{X_iX_i}^\# \overline{\mathbf{R}}_{X_iX_j} \mathbf{I}\} \rangle = \langle (\mathbf{I} \otimes \mathbf{R}_{X_iX_i}^{(q)\#}) \text{vec}\{\overline{\mathbf{R}}_{X_iX_j}\} \rangle$$

where the last equality is due to (48c). Since $\mathbf{x}(t)$ is independent of $\mathbf{x}(t')$ if $t \neq t'$ (Section III-A), the double sum merges into a single index q , leaving only

$$\begin{aligned} E\{\mathbf{g}_{ij}\mathbf{g}_{kl}^\dagger\} &= \frac{1}{T^2} \sum_{q=1}^Q n_q^2 (\mathbf{I} \otimes \mathbf{R}_{X_iX_i}^{(q)\#}) \\ &\quad \cdot E\{\text{vec}\{\overline{\mathbf{R}}_{X_iX_j}^{(q)}\} \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_kX_l}^{(q)}\}\} (\mathbf{I} \otimes \mathbf{R}_{X_kX_k}^{(q)\#}). \end{aligned} \quad (62)$$

In a second step, we work out the covariance matrix of the vectorized empirical matrices within (62) as follows. For $i \neq j$,

$$\begin{aligned}
& E\{\text{vec}\{\overline{\mathbf{R}}_{X_i X_j}^{(q)}\} \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_i X_j}^{(q)}\}\} \\
&= \frac{1}{n_q^2} \sum_{t \in \mathcal{D}_q} \sum_{r \in \mathcal{D}_q} E\{\text{vec}\{\mathbf{x}_i(t) \mathbf{x}_j^\dagger(t)\} \text{vec}^\dagger\{\mathbf{x}_i(r) \mathbf{x}_j^\dagger(r)\}\} \\
&= \frac{1}{n_q^2} \sum_{t \in \mathcal{D}_q} \sum_{r \in \mathcal{D}_q} E\{\mathbf{x}_j(t) \mathbf{x}_j^\dagger(r) \otimes \mathbf{x}_i(t) \mathbf{x}_i^\dagger(r)\} \\
&= \frac{1}{n_q^2} \sum_{t \in \mathcal{D}_q} \sum_{r \in \mathcal{D}_q} E\{\mathbf{x}_j(t) \mathbf{x}_j^\dagger(r)\} \otimes E\{\mathbf{x}_i(t) \mathbf{x}_i^\dagger(r)\} \\
&= \frac{1}{n_q^2} \sum_{t \in \mathcal{D}_q} \sum_{r \in \mathcal{D}_q} (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}) \delta_{rt} \\
&= \frac{1}{n_q} \mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}. \tag{63}
\end{aligned}$$

The first equality is an expansion of the definition of $\overline{\mathbf{R}}_{X_i X_j}^{(q)}$. The second equality uses Property D.1 below. The third equality is by independence of components i and j . The fourth equality is based on the assumption of Section III-A that $\forall t, r \in \mathcal{D}_q$, $E\{\mathbf{x}_i(t) \mathbf{x}_i^\dagger(r)\} = \mathbf{R}_{X_i X_i}^{(q)} \delta_{rt}$.

Substituting (63) into (62) with $k = i, l = j$,

$$\begin{aligned}
& E\{\mathbf{g}_{ij} \mathbf{g}_{ij}^\dagger\} \\
&= \frac{1}{T^2} \sum_{q=1}^Q \frac{n_q^2}{n_q} (\mathbf{I} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}) (\mathbf{I} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) \\
&= \frac{1}{T^2} \sum_{q=1}^Q n_q (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) = \frac{1}{T} \mathbf{H}_{ij} \tag{64}
\end{aligned}$$

which establishes (60). In (64), the second equality is due to (48a) and then (49b); the last step follows from (34). In order to establish (61), we right-multiply (63) by the commutation matrix $\mathcal{T}_{m,m}$, defined in Section IV-B, and use the equality $\text{vec}^\dagger\{\overline{\mathbf{R}}_{X_i X_j}^{(q)}\} \mathcal{T}_{m,m} = \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_j X_i}^{(q)}\}$, which follows from (35). This turns (63) into

$$\begin{aligned}
& E\{\text{vec}\{\overline{\mathbf{R}}_{X_i X_j}^{(q)}\} \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_j X_i}^{(q)}\}\} \\
&= \frac{1}{n_q} (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}) \mathcal{T}_{m,m}. \tag{65}
\end{aligned}$$

Substituting (65) in (62) with $k = j, l = i$,

$$\begin{aligned}
& E\{\mathbf{g}_{ij} \mathbf{g}_{ji}^\dagger\} = \frac{1}{T^2} \sum_{q=1}^Q \frac{n_q^2}{n_q} (\mathbf{I} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) \\
&\quad \cdot (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}) \mathcal{T}_{m,m} (\mathbf{I} \otimes \mathbf{R}_{X_j X_j}^{(q)\#}) \\
&= \frac{1}{T} (\mathbf{R}_{X_j X_j}^{(q)} \mathbf{R}_{X_j X_j}^{(q)\#} \otimes \mathbf{R}_{X_i X_i}^{(q)} \mathbf{R}_{X_i X_i}^{(q)\#}) \mathcal{T}_{m,m} \\
&= \frac{1}{T} (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*) \mathcal{T}_{m,m} = \frac{1}{T} (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*) \mathcal{T}_{m,m} \tag{66}
\end{aligned}$$

where in the second equality we applied the property [41, Theorem 3.1] that for any two matrices $\mathbf{M}_{m \times n}$ and $\mathbf{N}_{p \times q}$,

$$\mathcal{T}_{m,p} (\mathbf{N} \otimes \mathbf{M}) = (\mathbf{M} \otimes \mathbf{N}) \mathcal{T}_{n,q}, \tag{67}$$

followed by (48a) and (15). In the third equality of (66), we applied (57). Finally,

$$\begin{aligned}
(\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*) \mathcal{T}_{m,m} &= (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*)^2 \mathcal{T}_{m,m} \tag{68} \\
&= (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*) \mathcal{T}_{m,m} (\mathbf{\Pi}_i^* \otimes \mathbf{\Pi}_j^*) = \mathcal{T}_{m,m}^{\mathbf{\Pi}_{ij}}
\end{aligned}$$

where the second equality is due to (67), and the last step is by (40). Substituting (68) in (66) concludes the proof of (61).

Property D.1: For any four vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$,

$$\text{vec}\{\mathbf{a} \mathbf{b}^\dagger\} \text{vec}^\dagger\{\mathbf{c} \mathbf{d}^\dagger\} = \mathbf{d} \mathbf{b}^\dagger \otimes \mathbf{a} \mathbf{c}^\dagger. \tag{69}$$

Proof: Equation (69) follows from

$$\begin{aligned}
\text{vec}\{\mathbf{a} \mathbf{b}^\dagger\} \text{vec}^\dagger\{\mathbf{c} \mathbf{d}^\dagger\} &= (\mathbf{b} \otimes \mathbf{a}) (\mathbf{d} \otimes \mathbf{c})^\dagger \tag{70} \\
&= (\mathbf{b} \otimes \mathbf{a}) (\mathbf{d}^\dagger \otimes \mathbf{c}^\dagger) = \mathbf{b} \mathbf{d}^\dagger \otimes \mathbf{a} \mathbf{c}^\dagger.
\end{aligned}$$

The first equality of (70) is based on the property that for any two vectors \mathbf{a} and \mathbf{b} , $\text{vec}\{\mathbf{a} \mathbf{b}^\dagger\} = \text{vec}\{\mathbf{a} \cdot \mathbf{1} \cdot \mathbf{b}^\dagger\} = \mathbf{b} \otimes \mathbf{a}$, where the second step is due to (48c). The second equality of (70) is based on (48b) and the third on (48a). \square

APPENDIX E FINAL FORM OF (41)

In this appendix we prove the second equality in (41). The desired identity is

$$\mathcal{H}^{-1} \mathcal{H}_{\mathbf{\Pi}} \mathcal{H}^{-\dagger} = \mathcal{H}_{\mathbf{\Pi}}^\# \tag{71}$$

Proof: Denote $\mathbf{M} = \mathcal{H}_{\mathbf{\Pi}}$, $\mathbf{N} = \mathcal{H} - \mathcal{H}_{\mathbf{\Pi}}$ and $\mathbf{\Pi} = \begin{bmatrix} \mathbf{\Pi}_j \otimes \mathbf{\Pi}_i & \mathbf{0}_{m^2 \times m^2} \\ \mathbf{0}_{m^2 \times m^2} & \mathbf{\Pi}_i \otimes \mathbf{\Pi}_j \end{bmatrix}$. With these notations and the invertibility of \mathcal{H} (which is assured if the identifiability conditions in Section V hold), the orthogonality conditions $\mathbf{\Pi} \mathbf{M} \mathbf{\Pi} = \mathbf{M}$ and $(\mathbf{I} - \mathbf{\Pi}) \mathbf{N} (\mathbf{I} - \mathbf{\Pi}) = \mathbf{N}$, which are required to fulfill Property E.1 below, hold, as can be readily verified. The equality in (71) thus results from Property E.1. \square

Property E.1: Let \mathbf{M} and \mathbf{N} be two symmetric matrices so that $\mathbf{M} + \mathbf{N}$ is invertible. Assume that \mathbf{M} and \mathbf{N} have orthogonal range spaces. Then

$$(\mathbf{M} + \mathbf{N})^{-1} \mathbf{M} (\mathbf{M} + \mathbf{N})^{-\dagger} = \mathbf{M}^\# \tag{72}$$

Proof: The outline of the proof is as follows. Multiply both sides of (72) with $(\mathbf{M} + \mathbf{N})$. Applying the four criteria of the Moore–Penrose pseudoinverse (49) to the resulting four summands on the RHS, three zero out using the orthogonal range spaces of \mathbf{M} and \mathbf{N} and the fact that $\mathbf{M} \mathbf{M}^\#$ and $\mathbf{M}^\# \mathbf{M}$ are orthogonal projection operators. The only nonzero term is due to (49a). \square

APPENDIX F EXPLICIT FORM FOR $\text{Cov}(\text{vec}\{\mathcal{E}_{ij}\})$ AND $\text{Cov}(\text{vec}\{\mathcal{E}_{ij}\}, \text{vec}\{\mathcal{E}_{ji}\})$

In this appendix, we obtain an explicit expression for the blocks of matrix $\mathcal{H}_{\mathbf{\Pi}}^\#$, where $\mathcal{H}_{\mathbf{\Pi}}$ is defined in (39). This calculation is the final step in the derivation of $\text{Cov}(\text{vec}\{\mathcal{E}_{ij}\})$ and $\text{Cov}(\text{vec}\{\mathcal{E}_{ij}\}, \text{vec}\{\mathcal{E}_{ji}\})$, given in (42) and (43).

In order to decompose \mathcal{H}_Π and \mathcal{H}_Π^\sharp into more basic terms, we introduce the decomposition $\mathbf{R}_{X_i X_i}^{(q)} = \mathbf{U}_i \Upsilon_i^{(q)} \mathbf{U}_i^\dagger$ where, for convenience, the $m_i \times m_i$ invertible matrices $\Upsilon_i^{(q)}$ are chosen so that the $m \times m_i$ matrices \mathbf{U}_i have orthonormal columns. \mathbf{U}_i are thus the orthonormalized \mathbf{A}_i of (4). It then holds that $\mathbf{U}_i^\dagger \mathbf{U}_i = \mathbf{I}$,

$$\mathbf{U}_i^\sharp = \mathbf{U}_i^\dagger, \quad (73)$$

and $\mathbf{R}_{X_i X_i}^{\sharp(q)} = \mathbf{U}_i \Upsilon_i^{-1(q)} \mathbf{U}_i^\dagger$, due to (24). With the notations $\mathcal{I} \triangleq \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathcal{T}_{m_i, m_j} \end{bmatrix}$, $\mathbf{U} \triangleq \begin{bmatrix} \mathbf{U}_j \otimes \mathbf{U}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_i \otimes \mathbf{U}_j \end{bmatrix}$ and $\mathcal{H}_\Upsilon \triangleq \begin{bmatrix} \langle \Upsilon_j \otimes \Upsilon_i^{-1} \rangle & \mathbf{I} \\ \mathbf{I} & \langle \Upsilon_j^{-1} \otimes \Upsilon_i \rangle \end{bmatrix}$, we can rewrite

$$\mathcal{H}_\Pi = \mathbf{U} \mathcal{I} \mathcal{H}_\Upsilon \mathcal{I}^\dagger \mathbf{U}^\dagger. \quad (74)$$

Then, if \mathcal{H}_Υ is indeed invertible (see discussion in Section V),

$$\mathcal{H}_\Pi^\sharp = \mathbf{U} \mathcal{I}^{-1} \mathcal{H}_\Upsilon^{-1} \mathcal{I}^{-\dagger} \mathbf{U}^\dagger \quad (75)$$

which follows from (24), (73), and the fact that \mathcal{I} is always a full-rank matrix. Since \mathbf{U} and \mathcal{I} are block-diagonal matrices, then in order to calculate the blocks of \mathcal{H}_Π^\sharp , all that remains is to invert \mathcal{H}_Υ .

For the upper-left $m_i m_j \times m_i m_j$ block of $\mathcal{H}_\Upsilon^{-1}$, a block-matrix inversion formula [47] yields

$$[\mathcal{H}_\Upsilon^{-1}]^{\text{UL}} = (\langle \Upsilon_j \otimes \Upsilon_i^{-1} \rangle - \langle \Upsilon_j^{-1} \otimes \Upsilon_i \rangle^{-1})^{-1}. \quad (76)$$

Multiplying both sides by the upper-left blocks of \mathbf{U} and \mathcal{I}^{-1} ,

$$\begin{aligned} [\mathcal{H}_\Pi^\sharp]^{\text{UL}} &= (\mathbf{U}_j \otimes \mathbf{U}_i) (\langle \Upsilon_j \otimes \Upsilon_i^{-1} \rangle - \langle \Upsilon_j^{-1} \otimes \Upsilon_i \rangle^{-1})^{-1} \\ &\quad \cdot (\mathbf{U}_j \otimes \mathbf{U}_i)^\dagger \\ &= (\langle \mathbf{R}_{X_j X_j} \otimes \mathbf{R}_{X_i X_i}^\sharp \rangle - \langle \mathbf{R}_{X_j X_j}^\sharp \otimes \mathbf{R}_{X_i X_i} \rangle)^\sharp \end{aligned} \quad (77)$$

which follows from (24), (73), and (75). Similar arguments yield the expression

$$\begin{aligned} [\mathcal{H}_\Pi^\sharp]^{\text{UR}} &= -(\mathbf{U}_j \otimes \mathbf{U}_i) (\langle \Upsilon_j \otimes \Upsilon_i^{-1} \rangle - \langle \Upsilon_j^{-1} \otimes \Upsilon_i \rangle^{-1})^{-1} \\ &\quad \cdot \langle \Upsilon_j^{-1} \otimes \Upsilon_i \rangle^{-1} \mathcal{T}_{m_i, m_j} (\mathbf{U}_i \otimes \mathbf{U}_j)^\dagger \end{aligned} \quad (78)$$

for the upper-right $m_i m_j \times m_i m_j$ block of \mathcal{H}_Π^\sharp . Inserting the neutral term $(\mathbf{U}_j^\dagger \otimes \mathbf{U}_i^\dagger)(\mathbf{U}_j \otimes \mathbf{U}_i) = \mathbf{I}$ at the ‘ \cdot ’ symbol in (78) and then applying algebraic operations similarly to (77),

$$\begin{aligned} [\mathcal{H}_\Pi^\sharp]^{\text{UR}} &= -(\langle \mathbf{R}_{X_j X_j} \otimes \mathbf{R}_{X_i X_i}^\sharp \rangle - \langle \mathbf{R}_{X_j X_j}^\sharp \otimes \mathbf{R}_{X_i X_i} \rangle)^\sharp \\ &\quad \cdot \langle \mathbf{R}_{X_j X_j}^\sharp \otimes \mathbf{R}_{X_i X_i} \rangle^\sharp \mathcal{T}_{m, m}. \end{aligned} \quad (79)$$

By exchanging i with j in (77) and (79), one obtains the two lower blocks of \mathcal{H}_Π^\sharp . Substituting these results in (41) yields (42) and (43).

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous reviewers for their helpful comments and suggestions.

- [1] J.-F. Cardoso, “Multidimensional independent component analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Seattle, WA, May 1998, vol. 4, pp. 1941–1944.
- [2] F. Kohl, G. Wübbeler, D. Kolossa, C. Elster, M. Bär, and R. Orglmeister, “Non-independent BSS: A model for evoked MEG signals with controllable dependencies,” in *Independent Component Analysis and Signal Separation*, T. Adalı, C. Jutten, J. M. T. Romano, and A. K. Barros, Eds. Heidelberg, Germany: Springer, 2009, pp. 443–450.
- [3] A. Ossadchi, S. Baillet, J. C. Mosher, D. Thyerlei, W. Sutherling, and R. M. Leahy, “Automated interictal spike detection and source localization in magnetoencephalography using independent components analysis and spatio-temporal clustering,” *Clin. Neurophysiol.*, vol. 115, no. 3, pp. 508–522, Mar. 2004.
- [4] 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.
- [5] L. De Lathauwer, B. De Moor, and J. Vandewalle, “Fetal electrocardiogram extraction by source subspace separation,” in *Proc. IEEE SP/ATHOS Workshop HOS*, Girona, Spain, Jun. 1995, pp. 134–138.
- [6] L. Bedini, D. Herranz, E. Salerno, C. Baccigalupi, E. E. Kuruoğlu, and A. Tonazzini, “Separation of correlated astrophysical sources using multiple-lag data covariance matrices,” in *EURASIP J. Adv. Signal Process.*, 2005, vol. 2005, no. 15, pp. 2400–2412.
- [7] 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.
- [8] D. Lahat, J.-F. Cardoso, M. Le Jeune, and H. Messer, “Multidimensional ICA and its performance analysis applied to CMB observations,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 22–27, 2011, pp. 3724–3727.
- [9] A. Hyvärinen, “Beyond independent components,” in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, Edinburgh, U.K., Sep. 1999, vol. 2, pp. 809–814.
- [10] A. Hyvärinen and U. Köster, “FastISA: A fast fixed-point algorithm for independent subspace analysis,” presented at the Eur. Symp. Artif. Neural Netw. (ESANN), Bruges, Belgium, Apr. 2006.
- [11] H. Shen, K. Hüper, and M. Kleinstüber, “Block Jacobi-type methods for log-likelihood based linear independent subspace analysis,” in *Proc. Mach. Learn. Signal Process. (MLSP)*, Thessaloniki, Greece, Aug. 2007, pp. 133–138.
- [12] H. Shen and K. Hüper, “Generalised FastICA for independent subspace analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Honolulu, HI, Apr. 2007, vol. 4, pp. IV-1409–IV-1412.
- [13] J.-F. Cardoso, “Blind signal separation: Statistical principles,” *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.
- [14] J.-F. Cardoso, “On the performance of orthogonal source separation algorithms,” in *Proc. EUSIPCO*, Edinburgh, U.K., Sep. 1994, pp. 776–779.
- [15] L. De Lathauwer, B. De Moor, and J. Vandewalle, “A prewhitening-induced bound on the identification error in independent component analysis,” *IEEE Trans. Circuits Syst. I*, vol. 52, no. 3, pp. 546–554, Mar. 2005.
- [16] F. J. Theis, “Towards a general independent subspace analysis,” in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, vol. 19, pp. 1361–1368.
- [17] H. Choi and S. Choi, “Relative gradient learning for independent subspace analysis,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Vancouver, Canada, Jul. 2006, pp. 3919–3924.
- [18] S. Kirshner and B. Póczos, “ICA and ISA using Schweizer-Wolff measure of dependence,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Helsinki, Finland, 2008, pp. 464–471.
- [19] P. Gruber, H. W. Gutch, and F. J. Theis, “Hierarchical extraction of independent subspaces of unknown dimensions,” in *Independent Component Analysis and Signal Separation*, T. Adalı, C. Jutten, J. M. T. Romano, and A. K. Barros, Eds. Heidelberg, Germany: Springer, 2009, pp. 259–266.
- [20] Z. Szabó, B. Póczos, and A. Lőrincz, “Undercomplete blind subspace deconvolution,” *J. Mach. Learn. Res.*, vol. 8, pp. 1063–1095, May 2007.
- [21] O. Shalvi and E. Weinstein, “Maximum likelihood and lower bounds in system identification with non-Gaussian inputs,” *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 328–339, Mar. 1994.

- [22] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique based on second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.
- [23] V. Vigneron and C. Jutten, "Fisher information in source separation problems," in *Independent Component Analysis and Blind Signal Separation*, ser. LNCS. Heidelberg, Germany: Springer, Sep. 2004, vol. 3195, pp. 168–176.
- [24] P. Tichavský, Z. Koldovský, and E. Oja, "Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis," *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1189–1203, Apr. 2006.
- [25] E. Ollila, H.-J. Kim, and V. Koivunen, "Compact Cramér-Rao bound expression for independent component analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1421–1428, Apr. 2008.
- [26] E. Doron, A. Yeredor, and P. Tichavský, "Cramér-Rao-induced bound for blind separation of stationary parametric Gaussian sources," *IEEE Signal Process. Lett.*, vol. 14, no. 6, pp. 417–420, Jun. 2007.
- [27] J.-F. Cardoso and D.-T. Pham, "Separation of non-stationary sources: Algorithms and performance," in *Independent Component Analysis: Principles and Practice*. New York: Cambridge Univ. Press, 2001, pp. 158–180.
- [28] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.
- [29] J.-F. Cardoso, "Separation of non stationary sources; achievable performance," in *Proc. IEEE 10th Workshop Stat. Signal, Array Process.*, Pocono Manor, PA, Aug. 2000, pp. 359–363.
- [30] 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. Heidelberg, Germany: Springer, 2012, vol. 7191, pp. 155–162.
- [31] D. Lahat, J.-F. Cardoso, and H. Messer, "Optimal performance of second-order multidimensional ICA," in *Independent Component Analysis and Signal Separation*, T. Adalı, C. Jutten, J. M. T. Romano, and A. K. Barros, Eds. Heidelberg, Germany: Springer, 2009, pp. 50–57.
- [32] 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.
- [33] J.-F. Cardoso, "The three easy routes to independent component analysis; contrasts and geometry," in *Proc. Independent Component Anal. (ICA)*, San Diego, CA, Dec. 2001, pp. 1–6.
- [34] D. Lahat, J.-F. Cardoso, and H. Messer, "ICA of correlated sources mis modeled as uncorrelated: Performance analysis," in *Proc. IEEE/SP 15th Workshop Stat. Signal Process.*, Aug. 31–Sep. 3, 2009, pp. 489–492.
- [35] H. W. Gutch and F. J. Theis, "Independent subspace analysis is unique, given irreducibility," in *Independent Component Analysis and Signal Separation*, ser. LNCS, M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumley, Eds. Heidelberg, Germany: Springer, 2007, vol. 4666, pp. 49–56.
- [36] P. Comon, "Independent component analysis," in *Proc. Int. Signal Process. Workshop HOS*, Chamrousse, France, Jul. 1991, pp. 111–120.
- [37] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
- [38] H. Bousbia-Salah, A. Belouchrani, and K. Abed-Meraim, "Blind separation of non stationary sources using joint block diagonalization," in *Proc. Stat. Signal Process. (SSP)*, Aug. 2001, pp. 448–451.
- [39] D.-T. Pham, "Blind separation of cyclostationary sources using joint block approximate diagonalization," in *Independent Component Analysis and Signal Separation*, ser. LNCS. Heidelberg, Germany: Springer, 2007, vol. 4666, pp. 244–251.
- [40] D.-T. Pham, "Joint approximate diagonalization of positive definite hermitian matrices," *SIAM J. Matrix Anal. Appl.*, vol. 22, no. 4, pp. 1136–1152, 2001.
- [41] J. R. Magnus and H. Neudecker, "The commutation matrix: Some properties and applications," *Ann. Stat.*, vol. 7, no. 2, pp. 381–394, Mar. 1979.
- [42] R. E. Bellman, *Introduction to Matrix Analysis*. New York: McGraw-Hill, 1960.
- [43] 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.
- [44] L. Tong, R. Liu, V. Soon, and Y. F. Huang, "Indeterminacy and identifiability of blind identification," *IEEE Trans. Circuits Syst.*, vol. 38, no. 5, pp. 499–509, May 1991.
- [45] H. W. Gutch, T. Maehara, and F. J. Theis, "Second order subspace analysis and simple decompositions," in *Latent Variable Analysis and Signal Separation*, ser. LNCS, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, Eds. Heidelberg, Germany: Springer, 2010, vol. 6365, pp. 370–377.
- [46] A. Graham, *Kronecker Products and Matrix Calculus With Applications*, ser. Mathematics and its Applications. Chichester, West Sussex, U.K.: Ellis Horwood, 1981.
- [47] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Oct. 2008, version 20081110 [Online]. Available: <http://www2.imm.dtu.dk/pubdb/p.php?3274>
- [48] P. Comon, "Independent component analysis," in *HOS (Keynote Address)*, J.-L. Lacoume, Ed. New York: Elsevier, 1992, pp. 29–38, re-published.