

Gradient Algorithms for Complex Non-Gaussian Independent Component/Vector Extraction, Question of Convergence

Zbyněk Koldovský¹, Senior Member, IEEE, and Petr Tichavský², Senior Member, IEEE

Abstract—We revise the problem of extracting one independent component from an instantaneous linear mixture of signals. The mixing matrix is parameterized by two vectors: one column of the mixing matrix, and one row of the demixing matrix. The separation is based on the non-Gaussianity of the source of interest, while the remaining background signals are assumed to be Gaussian. Three gradient-based estimation algorithms are derived using the maximum likelihood principle and are compared with the Natural Gradient algorithm for Independent Component Analysis and with One-Unit FastICA based on negentropy maximization. The ideas and algorithms are also generalized to the extraction of a vector component when the extraction proceeds jointly from a set of instantaneous mixtures. Throughout this paper, we address the problem concerning the size of the region of convergence for which the algorithms guarantee the extraction of the desired source. We show that the size is influenced by the signal-to-interference ratio on the input channels. Simulations comparing several algorithms confirm this observation. They show a different size of the region of convergence under a scenario in which the source of interest is dominant or weak. Here, our proposed modifications of the gradient methods, taking into account the dominance/weakness of the source, show improved global convergence.

Index Terms—Blind source separation, blind source extraction, independent component analysis, independent vector analysis.

I. INTRODUCTION

A. Independent Component Analysis

INDEPENDENT Component Analysis (ICA) has been a popular method proposed for Blind Source Separation (BSS) since the 1990s [1]–[4]. In the basic ICA model, signals observed on d sensors are assumed to be linear mixtures of

Manuscript received January 22, 2018; revised June 25, 2018, October 15, 2018, and November 21, 2018; accepted December 9, 2018. Date of publication December 17, 2018; date of current version January 4, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wenwu Wang. This paper was presented at the 25th European Signal Processing Conference, Aug. 28–Sep. 2, 2017, in Kos, Greece. This work was supported in part by The Czech Science Foundation through Project 17-00902S and in part by the United States Department of the Navy, Office of Naval Research Global, through Project N62909-18-1-2040. (Corresponding author: Zbyněk Koldovský.)

Z. Koldovský is with the Acoustic Signal Analysis and Processing Group, Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec, Liberec 46117, Czech Republic (e-mail: zbynek.koldovsky@tul.cz).

P. Tichavský is with The Czech Academy of Sciences, Institute of Information Theory and Automation, Praha 18208, Czech Republic (e-mail: tichavsk@utia.cas.cz).

Digital Object Identifier 10.1109/TSP.2018.2887185

d “original” signals, which are mutually independent in the statistical sense. The mixing model is given by

$$\mathbf{x}(n) = \mathbf{A}\mathbf{u}(n), \quad (1)$$

where $\mathbf{x}(n)$ is a $d \times 1$ vector of the mixed signals; \mathbf{A} is a $d \times d$ nonsingular mixing matrix; $\mathbf{u}(n)$ is a $d \times 1$ vector of the original signals; and n denotes the sample index. In the non-Gaussianity-based ICA, the j th original signal $u_j(n)$ (the j th element of $\mathbf{u}(n)$) is modeled as an independently and identically distributed (i.i.d.) sequence of random variables with the probability density function (pdf) $p_j(\cdot)$. The goal is to estimate \mathbf{A}^{-1} using $\mathbf{x}(n)$, $n = 1, \dots, N$, through finding a square de-mixing matrix \mathbf{W} such that $\mathbf{W}\mathbf{x}(n)$ are independent or as close to independent as possible. In the discussion that follows, we will omit the sample index n for the sake of brevity, except where it is required.

While our focus in this paper is on complex-valued signals and parameters, our conclusions are valid for the real-valued case as well.

B. Blind Extraction of One Independent Source

This work addresses the problem of extraction (separation) of one independent component, which is often sufficient in applications such as speaker source enhancement, passive radar and sonar, or in biomedical signal processing. The complete decomposition performed by ICA can be computationally very demanding and superfluous. This is especially remarkable when there is a large number of sensors (say, 10 or more), or when a large number of mixtures (say, 128 or more) are separated in parallel, as in the Frequency-Domain ICA (FD-ICA) [5]. The idea of extracting only one source can also be applied in joint BSS [6]–[8], especially in Independent Vector Analysis (IVA) [9]. Here, the “source” is represented by a vector of separated components from the mixtures that are mutually dependent (but independent of the other vector components).

BSS involves the indeterminacy of order of the original signals in the mixture [2], [10], and therefore certain partial knowledge about the source to be extracted must be available to determine which independent component is the one of our interest. For example, the prior knowledge could be an expected direction of arrival (DOA) of the source, location of the source within a confined area [11], a property such as dominance within an angular range [12] or temporal structure [13], and so forth. Throughout this paper, we will assume that such knowledge

is available in the form of an initial value of the (de)mixing parameter. The wanted signal will be referred to as *source of interest (SOI)* while the rest of the mixture will be referred to as *background*.

The theoretical part of this paper will be constrained to the determined case, which means that the background is assumed to be a mixture of $d - 1$ latent variables, or, in other words, the whole mixture obeys the determined mixing as in (1). This assumption need not be overly restrictive when only one source should be extracted. Indeed, when the mixture \mathbf{x} consists of more than d sources (underdetermined case), algorithms based on the determined model can still be used provided that they are sufficiently robust against mismodeling and noise. However, these issues would go beyond the scope of this paper.

C. State-of-the-Art

The blind separation of one particular non-Gaussian source has already been studied in several contexts and some authors refer to it as Blind Signal Extraction (BSE) [13]–[15]. Projection pursuit [16], a technique used for exploratory data analysis, aims at finding “interesting” projections, including 1-D signals. This “interestingness” is defined through various measures reflecting the distance of the projected signal’s pdf from the Gaussian distribution [2], [17]. Various criteria of the interestingness were also derived in other contexts. For example, kurtosis appears in methods for blind adaptive beamforming or as a higher-order cumulant-based contrast; see, e.g., [18]–[20].

This framework was unified under ICA based on information theory [21]. Namely, the independence of signals can be measured using mutual information, which is the Kullback-Leibler divergence between the joint density of signals and the product of their marginal densities. The signals are independent if and only if that mutual information equals zero. Provided that the elements of $\mathbf{y}(n)$ are not correlated, the mutual information of $\mathbf{y}(n)$ is equal to the sum of entropies of $y_1(n), \dots, y_d(n)$, up to a constant. Hence, it follows that an independent component can be sought through minimizing the entropy of the separated signal under the constraint $E[|\mathbf{w}^H \mathbf{x}|^2] = 1$; here $E[\cdot]$ stands for the expectation operator; and \mathbf{w}^H is a de-mixing vector (a row of the de-mixing matrix \mathbf{W}). The fact that entropy is a measure of non-Gaussianity reveals the connection between the ICA-based separation of one signal and the contrast-based BSE techniques [2], [22].

In fact, many ICA methods apply d BSE estimators sequentially [23] or in parallel [24] to find all independent components in the mixture. The orthogonal constraint, which requires that the sample correlation between separated signals be equal to zero, is imposed on the BSE outputs in order to prevent the algorithms from finding any components twice. For example, the well-known FastICA algorithm has three basic variants: One-Unit FastICA is a BSE method optimizing the component’s non-Gaussianity [25]; Deflation FastICA applies the one-unit version sequentially [26]; Symmetric FastICA runs d one-unit algorithms simultaneously [26], [27].

The separation accuracy of the above methods is known to be limited [28]. One-unit FastICA exploits only the non-

Gaussianity of SOI and does not use the non-Gaussianity of the background [29]. The accuracy levels of deflation and symmetric FastICA are limited due to the orthogonal constraint [30]. While the latter limitation can be overcome, as shown, e.g., in [31], the limited accuracy of the one-unit approach poses an open problem, unless the BSE is done through the complete ICA. By comparing the performance analyses from [28], [29], [32] and the Cramér-Rao bound for ICA [31], it follows that one-unit methods can approach the optimum performance only when the background is Gaussian, but not otherwise.

D. Contribution

In this paper, we revisit the BSE problem by considering it explicitly: the goal is to extract one component from the instantaneous mixture that is as close to being independent of the background as possible; we refer to this approach as *Independent Component Extraction (ICE)*. A re-parameterization of the mixing model is introduced, in which the number of parameters is minimal for the BSE problem (the mixing and the separating vectors related to the SOI). Then, a statistical model is adopted from ICA where the background is assumed to be jointly Gaussian. The classical maximum likelihood estimation of the mixing parameters is considered, by which simplistic gradient-based estimation algorithms are derived.¹ The ICE approach provides a deeper insight into the BSE problem. In particular, it points out the role of the orthogonal constraint and the fact that the constraint is inherently applied within One-Unit FastICA. This approach also reveals the role of the model of the background’s pdf.

It is worth pointing out here that similar mixing and statistical models have been considered in the methods that were designed for Cosmic Microwave Background extraction from the Wilkinson Microwave Anisotropy Probe (WMAP) data or from the more recent Planck Mission data [34]. However, there is one important difference, namely, the mixing vector related to the SOI is assumed to be known. These methods are known under the name of Internal Linear Combination (ILC) [35], [36]; see also [37].

For the practical output of this paper, we focus on the ability of BSE algorithms to ensure that the desired SOI is being extracted, not a different source.² This is a crucial aspect in BSE, which has little been studied previously. When the extraction of the SOI is not guaranteed, it is necessary to extract all sources and to find the desired one afterwards; this way, however, the advantage of performing only one BSE task is lost. Another motivation is that the permutation ambiguity can impair on-line separation. For example, a sudden change in the region of convergence (ROC) due to dynamic signals and/or mixing conditions can cause the current mixing vector estimate to occur within the ROC of a different source. The given algorithm then performs several

¹In this paper, an algorithm estimating the separating vector is introduced, compared to [33], where only the variant for estimating the mixing vector is described.

²We do not focus on the algorithms’ accuracy as this is already a well-studied problem. The accuracy of BSE methods is fundamentally limited by the Cramér-Rao bound, which is asymptotically attainable, e.g., by One-Unit FastICA [28]; cf. the last paragraph in Section I-C; see also [38].

“diverging” steps, during which the separated sources are being permuted; the separation is poor in between. Therefore, the size of the ROC is studied in this paper. Let us emphasize the fact that the ROC is algorithm-dependent and is highly influenced by the Signal-to-Interference ratio in the channels of the observed mixture. Our experiments show that the ROC can depend on whether the optimization proceeds in the mixing or de-mixing parameters. Based on this fact, we propose novel variants for the gradient algorithms where the optimization parameters are selected automatically.

Next, our ideas are generalized to the extraction of a vector component, the so-called *Independent Vector Extraction (IVE)*. Here, the problem defines several instantaneous mixtures to be treated simultaneously, using joint statistical models. The goal is to extract one independent component per mixture where the extracted components should be as dependent as possible. IVE is an extension of ICE similar to that of ICA to IVA [39], [40]. A gradient algorithm with the automatic selection of the optimization parameters is derived, similarly to that utilized in ICE. Our experiments show that the convergence of the proposed IVE algorithm is superior to that of ICE because the improved convergence within several mixtures has a positive influence on the convergence within the other mixtures; the effect of the automatic selection is thus multiplied.

The rest of this paper is organized as follows. Section II introduces algebraic and statistical models for ICE. Section III is devoted to gradient-based ICE algorithms. The ideas and algorithms are generalized to the extraction of vector components in Section IV. Section V presents results of simulations, and Section VI concludes the paper.

II. INDEPENDENT COMPONENT EXTRACTION

Nomenclature: The following notation will be used throughout the article. Plain letters denote scalars, bold lower-case letters denote vectors, and bold capital letters denote matrices. The Matlab conventions for matrix/vector concatenation and indexing will be used, e.g., $[1; \mathbf{g}] = [1, \mathbf{g}^T]^T$, and $(\mathbf{A})_{j,:}$ is the j th row of \mathbf{A} . Next, \mathbf{g}^T , $\bar{\mathbf{g}}$, and \mathbf{g}^H denote the transpose, the complex conjugate value and the conjugate transpose of \mathbf{g} , respectively. Symbolic scalar and vector random variables will be denoted by lower-case letters, e.g., s and \mathbf{x} , while the quantities collecting their N samples will be denoted by bold (capital) letters, e.g., \mathbf{s} and \mathbf{X} . Estimated values of signals will be denoted by hats, e.g., \hat{s} . For simplicity, the hat will be omitted in the case of estimated values of parameters, e.g., \mathbf{w} , unless it is necessary to distinguish between its estimated and true values.

A. Mixing Model Parameterization

Without any loss of generality, let the SOI be $s = u_1$ and \mathbf{a} be the first column of \mathbf{A} , so it can be partitioned as $\mathbf{A} = [\mathbf{a}, \mathbf{A}_2]$. Then, \mathbf{x} can be written in the form

$$\mathbf{x} = \mathbf{a}s + \mathbf{y}, \quad (2)$$

where $\mathbf{y} = \mathbf{A}_2\mathbf{u}_2$ and $\mathbf{u}_2 = [u_2, \dots, u_d]^T$. The single-target description (2) has been widely studied in array processing literature [41]. Here, the fact that $\mathbf{y} = \mathbf{A}_2\mathbf{u}_2$ means that we restrain

our considerations to the determined scenario (the mixture consists of the same number of sources as that of the sensors).

Let the new mixing matrix for ICE and its inverse matrix be denoted by \mathbf{A}_{ICE} and \mathbf{W}_{ICE} , respectively. In ICE, the identification of \mathbf{A}_2 or the decomposition of \mathbf{y} into independent signals is *not* the goal. Therefore, the structure of the mixing matrix is $\mathbf{A}_{\text{ICE}} = [\mathbf{a}, \mathbf{Q}]$ where \mathbf{Q} is, for now, arbitrary.

Then, (2) can be written as

$$\mathbf{x} = \mathbf{A}_{\text{ICE}}\mathbf{v}, \quad (3)$$

where $\mathbf{v} = [s; \mathbf{z}]$, and $\mathbf{y} = \mathbf{Q}\mathbf{z}$. It holds that \mathbf{z} spans the same subspace as that spanned by \mathbf{u}_2 .

To complete the mixing model definition, we look at the inverse matrix $\mathbf{W}_{\text{ICE}} = \mathbf{A}_{\text{ICE}}^{-1}$. Let \mathbf{a} and \mathbf{W}_{ICE} be partitioned, respectively, as

$$\mathbf{a} = \begin{pmatrix} \gamma \\ \mathbf{g} \end{pmatrix} \quad (4)$$

and

$$\mathbf{W}_{\text{ICE}} = \begin{pmatrix} \mathbf{w}^H \\ \mathbf{B} \end{pmatrix}. \quad (5)$$

\mathbf{B} is required to be orthogonal to \mathbf{a} , i.e., $\mathbf{B}\mathbf{a} = \mathbf{0}$, which ensures that the signals separated by the lower part of \mathbf{W}_{ICE} , namely, by $\mathbf{B}\mathbf{x}$, do not contain any contribution of s . A useful selection is

$$\mathbf{B} = (\mathbf{g} - \gamma\mathbf{I}_{d-1}), \quad (6)$$

where \mathbf{I}_d denotes the $d \times d$ identity matrix. Let \mathbf{w} be partitioned as

$$\mathbf{w} = \begin{pmatrix} \beta \\ \mathbf{h} \end{pmatrix}. \quad (7)$$

The de-mixing matrix then has the structure

$$\mathbf{W}_{\text{ICE}} = \begin{pmatrix} \mathbf{w}^H \\ \mathbf{B} \end{pmatrix} = \begin{pmatrix} \bar{\beta} & \mathbf{h}^H \\ \mathbf{g} & -\gamma\mathbf{I}_{d-1} \end{pmatrix}, \quad (8)$$

and from $\mathbf{A}_{\text{ICE}}^{-1} = \mathbf{W}_{\text{ICE}}$ it follows that

$$\mathbf{A}_{\text{ICE}} = (\mathbf{a} \mathbf{Q}) = \begin{pmatrix} \gamma & \mathbf{h}^H \\ \mathbf{g} & \frac{1}{\gamma}(\mathbf{g}\mathbf{h}^H - \mathbf{I}_{d-1}) \end{pmatrix}, \quad (9)$$

where β and γ are linked through

$$\bar{\beta}\gamma = 1 - \mathbf{h}^H\mathbf{g}. \quad (10)$$

The latter equation can also be written in the form $\mathbf{w}^H\mathbf{a} = 1$, which is known as the *distortionless response* constraint; see page 515 in [41]. The parameterization of the mixing and de-mixing matrices is similar to the one used in ILC [36], [37].

It is worth mentioning here that ICE and Multidimensional ICA [42] are similar to each other; the latter is also known as Independent Subspace Analysis (ISA) [43]. In ISA, the goal is to separate subspaces of components that are mutually independent while components inside of the subspaces can be dependent. The goal to separate one independent component thus could be formulated as a special case of ISA where \mathbf{u} is divided into two subspaces of dimensions 1 and $d - 1$, respectively. What makes ICE different is that the separation of the background subspace

from the SOI is not as ambiguous as in ISA, which is ensured by the structure of the de-mixing matrix.³

The free variables in the ICE mixing model are the elements of \mathbf{g} and \mathbf{h} , and one of the parameters β or γ . In total, there are $2d - 1$ free (real or complex) parameters. The role of \mathbf{a} , as follows from (2), is the *mixing vector* related to s , which is in beamforming literature also sometimes referred to as the steering vector. Next, \mathbf{w} is the *separating vector* as $s = \mathbf{w}^H \mathbf{x}$. For the background signal \mathbf{z} , it holds that

$$\mathbf{z} = \mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{A}_2\mathbf{u}_2. \quad (11)$$

Note that \mathbf{A}_2 is not identified in the model, so the relationship between \mathbf{z} and \mathbf{u}_2 remains unknown after performing ICE. The components of \mathbf{z} are independent after the extraction only when $d = 2$ or, for $d > 2$, in very special cases that $\mathbf{B}\mathbf{A}_2 = \mathbf{\Lambda}\mathbf{P}$; $\mathbf{\Lambda}$ denotes a diagonal (scaling) matrix with non-zero elements on the main diagonal, and \mathbf{P} denotes a permutation matrix.

B. Indeterminacies

Similarly to ICA, the scales of s and of \mathbf{a} are ambiguous in the sense that, in (2) they can be replaced, respectively, by αs and $\alpha^{-1} \mathbf{a}$ where $\alpha \neq 0$. The scaling ambiguity can be avoided by fixing β or γ . A specific case occurs when $\gamma = 1$ [33], since then s corresponds to the so-called spatial image of the SOI on the first sensor [44]–[46]. This can be useful for modeling the pdf of s , as the physical meaning of that scale is often known. By contrast, when no such knowledge is given, it might be better to keep γ (or β) free.

The second ambiguity is that the role of $s = u_1$ is interchangeable with any independent component of \mathbf{x} , that is, with any u_i , $i = 2, \dots, d$. This fact follows from the indeterminacy of the order of original components in the mixture [22], which causes the permutation problem in joint BSS; see, e.g., [47]. As was already stated, in this work we assume that an initial guess of either \mathbf{a} or \mathbf{w} is given.

C. Statistical Model

The main principle of ICE is the same as that of ICA. We make the assumption that s and \mathbf{z} are *independent*, and ICE is formulated as follows:

Find vectors \mathbf{a} and \mathbf{w} such that $\mathbf{w}^T \mathbf{x}$ and $\mathbf{B}\mathbf{x}$ are independent (or as close to independent as possible).

Let the pdf of s and of \mathbf{z} be, respectively, denoted by $p_s(\xi_1)$ and $p_z(\xi_2)$; ξ_1 and ξ_2 denote free variables of appropriate dimensions. The joint pdf of s and \mathbf{z} is, owing to their mutual independence,

$$p_s(\xi) = p_s(\xi_1) \cdot p_z(\xi_2), \quad (12)$$

³The fact that the structure of the de-mixing matrix is fixed makes ICE different from ISA (i.e., it is a different approach to the same problem). We do not claim, though, that ICE and ISA necessarily lead to different results.

where $\xi = [\xi_1; \xi_2]$. From (8), the joint pdf of the mixed signals $\mathbf{x} = \mathbf{A}_{\text{ICE}} \mathbf{v}$ is

$$p_{\mathbf{x}}(\xi) = p_s(\mathbf{w}^H \xi) \cdot p_z(\mathbf{B}\xi) \cdot |\det \mathbf{W}_{\text{ICE}}|^2 \quad (13)$$

$$= p_s(\bar{\beta}\xi_1 + \mathbf{h}^H \xi_2) \cdot p_z(\xi_1 \mathbf{g} - \gamma \xi_2) \cdot |\gamma|^{2(d-2)}, \quad (14)$$

where the identity

$$\det \mathbf{W}_{\text{ICE}} = (-1)^{d-1} \gamma^{d-2} \quad (15)$$

$$= (-1)^{d-1} \beta^{-(d-2)} (1 - \mathbf{h}^H \mathbf{g})^{d-2}, \quad (16)$$

was used, which can easily be verified from (8) using (10).

The log-likelihood function of N signal samples depends on \mathbf{a} and \mathbf{w} ; hence it is

$$\mathcal{L}(\mathbf{a}, \mathbf{w}) = \frac{1}{N} \log \prod_{n=1}^N p_{\mathbf{x}}(\mathbf{a}, \mathbf{w} | \mathbf{x}(n)) \quad (17)$$

$$= \frac{1}{N} \sum_{n=1}^N \log p_s(\mathbf{w}^H \mathbf{x}(n)) + \frac{1}{N} \sum_{n=1}^N \log p_z(\mathbf{B}\mathbf{x}(n)) + (d-2) \log |\gamma|^2. \quad (18)$$

D. Gaussian Background

As previously explained in Section II-A, the background signals are highly probable to remain mixed after ICE, unless $d = 2$. This opens the problem of modeling the pdf of \mathbf{z} . A straightforward choice is that the components of \mathbf{z} have the circularly symmetric Gaussian distribution with zero mean and covariance \mathbf{C}_z , i.e., $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_z)$. This choice can be justified by the fact that the said components are mixed and correlated; moreover, from the Central Limit Theorem it follows that their distribution is close to Gaussian [22]. The covariance matrix $\mathbf{C}_z = \mathbb{E}[\mathbf{z}\mathbf{z}^H]$ is a nuisance parameter.

In this paper, we restrain our considerations to this Gaussian background model, noting that other choices are worthy of future investigation [48]. It is a simplifying assumption, which means that only second-order statistics of the background are used. This model should also work for non-Gaussian background signals as long as their second-order statistics exist. Now, (18) takes the form

$$\begin{aligned} \mathcal{L}(\mathbf{a}, \mathbf{w}) &= \frac{1}{N} \sum_{n=1}^N \log p_s(\mathbf{w}^H \mathbf{x}(n)) \\ &\quad - \frac{1}{N} \sum_{n=1}^N \mathbf{x}(n)^H \mathbf{B}^H \mathbf{C}_z^{-1} \mathbf{B} \mathbf{x}(n) + (d-2) \log |\gamma|^2 \\ &\quad - \log \det \mathbf{C}_z - d \log \pi. \end{aligned} \quad (19)$$

E. Orthogonal Constraint

By inspecting (18) and (19), it can be seen that the link between \mathbf{a} and \mathbf{w} , which are both related to the SOI, is rather “weak”. Indeed, the first term on the right-hand side of (18) depends purely on \mathbf{w} , while the second and the third terms depend purely on \mathbf{a} . The only link between \mathbf{a} and \mathbf{w} is thus expressed in (10). Consequently, the log-likelihood function can

have spurious maxima where \mathbf{a} and \mathbf{w} do not jointly correspond to the SOI.

Many ICA algorithms impose the orthogonal constraint (OG) [30], which decreases the number of unknown parameters in the mixing model. This constraint can be used to avoid spurious solutions in ICE and to stabilize the convergence of algorithms. Let now \mathbf{W}_{ICE} denote an ICE de-mixing matrix estimate and

$$\widehat{\mathbf{V}} = \begin{pmatrix} \widehat{\mathbf{s}} \\ \widehat{\mathbf{z}} \end{pmatrix} = \mathbf{W}_{\text{ICE}} \mathbf{X} \quad (20)$$

be the estimated de-mixed signals, that is, $\widehat{\mathbf{s}}$ be the $1 \times N$ row vector of samples of the extracted SOI, and $\widehat{\mathbf{z}}$ be the $(d-1) \times N$ matrix of samples of the background signals. The OG means that

$$\frac{1}{N} \widehat{\mathbf{s}} \cdot \widehat{\mathbf{z}}^H = \frac{1}{N} \mathbf{w}^H \mathbf{X} \mathbf{X}^H \mathbf{B}^H = \mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{B}^H = \mathbf{0}, \quad (21)$$

where $\widehat{\mathbf{C}}_x = \mathbf{X} \mathbf{X}^H / N$ is the sample-based estimate of $\mathbf{C}_x = \mathbb{E}[\mathbf{x} \mathbf{x}^H]$.

The OG introduces a link between \mathbf{a} and \mathbf{w} , so \mathbf{W}_{ICE} is a function of either \mathbf{a} or \mathbf{w} . The dependencies, whose derivations are given in Appendix A, are

$$\mathbf{w} = \frac{\widehat{\mathbf{C}}_x^{-1} \mathbf{a}}{\mathbf{a}^H \widehat{\mathbf{C}}_x^{-1} \mathbf{a}}, \quad (22)$$

when \mathbf{a} is the dependent variable, and

$$\mathbf{a} = \frac{\widehat{\mathbf{C}}_x \mathbf{w}}{\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}}, \quad (23)$$

when the dependent variable is \mathbf{w} .

Interestingly, the coupling (22) corresponds to the approximation of

$$\mathbf{w}_{\text{MPDR}}^H \mathbf{x} = \frac{\mathbf{a}^H \mathbf{C}_x^{-1}}{\mathbf{a}^H \mathbf{C}_x^{-1} \mathbf{a}} \mathbf{x}, \quad (24)$$

which is the minimum-power distortionless (MPDR) beamformer steered in the direction given by \mathbf{a} , the well-known optimum beamformer in array processing theory [41]; see also [33]. In Appendix B, it is shown that if \mathbf{a} is equal to its true value, then

$$\mathbf{w}_{\text{MPDR}}^H \mathbf{x} = s. \quad (25)$$

The advantage of (23) is that the computation of \mathbf{a} does not involve the inverse of $\widehat{\mathbf{C}}_x$.

III. GRADIENT-BASED ICE ALGORITHMS

In this section, we derive gradient ICE algorithms aiming at the maximum likelihood estimation through searching for the maximum of (19). Since p_s and \mathbf{C}_z in (19) are not known, we propose a contrast function replacing the true one where p_s and \mathbf{C}_z are approximated in a certain way. This is sometimes referred to as the quasi-maximum likelihood approach; see, e.g., [49].

A. Optimization in \mathbf{w}

For the optimization in \mathbf{w} , given the coupling (23), β is selected as a free variable while γ is dependent. Following (19) where the last two terms are neglected as they are constant, the contrast function is defined as

$$\mathcal{C}(\mathbf{a}, \mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \left\{ \log f(\mathbf{w}^H \mathbf{x}(n)) - \mathbf{x}(n)^H \mathbf{B}^H \mathbf{R} \mathbf{B} \mathbf{x}(n) \right\} + (d-2) \log |\gamma|^2, \quad (26)$$

where $f(\cdot)$ is the model pdf of the target signal (replacing p_s), and \mathbf{R} is a weighting positive definite matrix (replacing \mathbf{C}_z^{-1}).

Using the Wirtinger calculus [50], [51], we derive in Appendix C that the gradient of \mathcal{C} with respect to \mathbf{w}^H , under the coupling (23), equals

$$\begin{aligned} \left. \frac{\partial \mathcal{C}}{\partial \mathbf{w}^H} \right|_{\text{w.r.t. (23)}} &= -\frac{1}{N} \mathbf{X} \phi(\mathbf{w}^H \mathbf{X})^T + 2\mathbf{a} \operatorname{tr}(\mathbf{R} \widehat{\mathbf{C}}_z) \\ &- (\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w})^{-1} (\widehat{\mathbf{C}}_x \mathbf{E}^H \mathbf{R} \widehat{\mathbf{C}}_z \mathbf{h} - \operatorname{tr}(\mathbf{R} \widehat{\mathbf{C}}_x \mathbf{E}^H) \widehat{\mathbf{C}}_x \mathbf{e}_1) \\ &- 2(d-2)\mathbf{a} + \bar{\gamma}^{-1} (d-2) (\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w})^{-1} \widehat{\mathbf{C}}_x \mathbf{e}_1, \end{aligned} \quad (27)$$

where $\operatorname{tr}(\cdot)$ denotes the trace; $\mathbf{E} = [\mathbf{0} \ \mathbf{I}_{d-1}]$; \mathbf{e}_1 denotes the first column of \mathbf{I}_d ; and

$$\phi(\xi) = -\frac{\partial \log f(\xi)}{\partial \xi} \quad (28)$$

is the score function of the model pdf $f(\cdot)$.

Now, we put $\mathbf{R} = \widehat{\mathbf{C}}_z^{-1}$; this is a choice for which the derivative of (19) with respect to the unknown parameter \mathbf{C}_z is equal to zero. Then, the following identities can be applied in (27).

$$\operatorname{tr}(\widehat{\mathbf{C}}_z^{-1} \widehat{\mathbf{C}}_z) = \operatorname{tr}(\mathbf{I}_{d-1}) = d-1, \quad (29)$$

$$\mathbf{E}^H \mathbf{h} + \beta \mathbf{e}_1 = \mathbf{w}, \quad (30)$$

$$\begin{aligned} \widehat{\mathbf{C}}_z^{-1} \mathbf{B} \widehat{\mathbf{C}}_x &= \widehat{\mathbf{C}}_z^{-1} \mathbf{B} \mathbf{X} \mathbf{X}^H / N \\ &= \widehat{\mathbf{C}}_z^{-1} \widehat{\mathbf{z}} [\widehat{\mathbf{s}}^H \ \widehat{\mathbf{z}}^H] \mathbf{A}_{\text{ICE}}^H = \mathbf{E} \mathbf{A}_{\text{ICE}}^H \end{aligned} \quad (31)$$

$$\begin{aligned} \operatorname{tr}(\widehat{\mathbf{C}}_z^{-1} \mathbf{B} \widehat{\mathbf{C}}_x \mathbf{E}^H) &= \operatorname{tr}(\mathbf{E} \mathbf{A}_{\text{ICE}}^H \mathbf{E}^H) \\ &= \bar{\gamma}^{-1} \operatorname{tr}(\mathbf{h} \mathbf{g}^H - \mathbf{I}_{d-1}) \\ &= -\beta - (d-2) \bar{\gamma}^{-1}, \end{aligned} \quad (32)$$

where we used (20) and (21); (27) is now simplified to

$$\left. \frac{\partial \mathcal{C}}{\partial \mathbf{w}^H} \right|_{\text{w.r.t. (23)}} = \mathbf{a} - \frac{1}{N} \mathbf{X} \phi(\mathbf{w}^H \mathbf{X})^T. \quad (33)$$

In fact, $\mathbf{R} = \widehat{\mathbf{C}}_z^{-1}$ depends on the current value of \mathbf{w} since $\widehat{\mathbf{C}}_z = \mathbf{B} \widehat{\mathbf{C}}_x \mathbf{B}^H$. It means that, with any estimate of \mathbf{w} , the distribution of $\widehat{\mathbf{z}} = \mathbf{B} \mathbf{X}$ is assumed to be $\mathcal{CN}(\mathbf{0}, \widehat{\mathbf{C}}_z)$, which obviously introduces little (or no) information into the contrast function. $\widehat{\mathbf{C}}_z$ is close to the true covariance \mathbf{C}_z only when \mathbf{a} is close to its true value.

For $N \rightarrow +\infty$, (33) takes on the form

$$\frac{\partial \mathcal{C}}{\partial \mathbf{w}^H} = \mathbf{a} - \mathbb{E}[\mathbf{x} \phi(\mathbf{w}^H \mathbf{x})]. \quad (34)$$

Algorithm 1: OGICE_w: Separating Vector Estimation Based on Orthogonally Constrained Gradient-ascent Algorithm.

Input: \mathbf{X} , \mathbf{w}_{ini} , μ , tol
Output: \mathbf{a} , \mathbf{w}

- 1 $\hat{\mathbf{C}}_{\mathbf{x}} = \mathbf{X}\mathbf{X}^H/N$;
- 2 $\mathbf{w} = \mathbf{w}_{\text{ini}}$;
- 3 **repeat**
- 4 $\lambda_{\mathbf{w}} \leftarrow (\mathbf{w}^H \hat{\mathbf{C}}_{\mathbf{x}} \mathbf{w})^{-1}$;
- 5 $\mathbf{a} \leftarrow \lambda_{\mathbf{w}} \hat{\mathbf{C}}_{\mathbf{x}} \mathbf{w}$; /* OG constraint (23) */
- 6 $\hat{\mathbf{s}} \leftarrow \mathbf{w}^H \mathbf{X}$;
- 7 $\nu \leftarrow \hat{\mathbf{s}} \phi(\hat{\mathbf{s}})^T / N$; /* due to cond. (36) */
- 8 $\Delta \leftarrow \mathbf{a} - \nu^{-1} \mathbf{X} \phi(\hat{\mathbf{s}})^T / N$; /* by (33) */
- 9 $\mathbf{w} \leftarrow \mathbf{w} + \mu \Delta$; /* gradient ascent */
- 10 **until** $\|\Delta\| < \text{tol}$;

When \mathbf{w} is the ideal separating vector, that is, when $\mathbf{w}^H \mathbf{x} = s$, then from (2) it follows that

$$\frac{\partial \mathcal{C}}{\partial \mathbf{w}^H} = (1 - \mathbb{E}[s\phi(s)])\mathbf{a}. \quad (35)$$

This shows us that the true separating vector is a stationary point of the contrast function only if $\phi(\cdot)$ satisfies the condition

$$\mathbb{E}[s\phi(s)] = 1. \quad (36)$$

Based on this observation, we propose a method whose steps are described in Algorithm 1. In every step, it iterates in the direction of the steepest ascent of \mathcal{C} where $\mathbf{R} = \hat{\mathbf{C}}_{\mathbf{z}}^{-1}$ (step 9), and $\phi(\cdot)$ is normalized so that condition (36) is satisfied for the current target signal estimate $\hat{\mathbf{s}} = \mathbf{w}^H \mathbf{x}$, that is, $\phi(\hat{\mathbf{s}}) \leftarrow \phi(\hat{\mathbf{s}})/(\hat{\mathbf{s}}\phi(\hat{\mathbf{s}})^T/N)$ (see steps 7 and 8). This is repeated until the norm of the gradient is smaller than tol ; μ is the step length parameter; and \mathbf{w}_{ini} is the initial guess. We call this method OGICE_w.

In fact, (34) coincides with the gradient of a heuristic criterion derived from mutual information in [32] (page 870, Eq. 4). The author, D.-T. Pham, called this approach ‘‘Blind Partial Separation’’. Our derivation provides a deeper insight into this result by showing its connection with maximum likelihood estimation. Most importantly, it is seen that (33) follows from a particular parameterization of the (de)-mixing model, it imposes the OG between \mathbf{a} and \mathbf{w} , and it relies on the Gaussian modeling of the background signals whose covariance is estimated as $\hat{\mathbf{C}}_{\mathbf{z}}$.

B. Optimization in \mathbf{a}

The gradient with respect to \mathbf{a} when \mathbf{w} is dependent through (22) and when $\gamma = 1$ has been derived in [33]. Treating γ as a free variable, and by putting $\mathbf{R} = \hat{\mathbf{C}}_{\mathbf{z}}^{-1}$, the gradient reads

$$\frac{\partial \mathcal{C}}{\partial \mathbf{a}^H} \Big|_{\text{w.r.t. (22)}} = \mathbf{w} - \frac{\lambda_{\mathbf{a}}}{N} \hat{\mathbf{C}}_{\mathbf{x}}^{-1} \mathbf{X} \phi(\mathbf{w}^H \mathbf{X})^T, \quad (37)$$

where $\lambda_{\mathbf{a}} = (\mathbf{a}^H \hat{\mathbf{C}}_{\mathbf{x}}^{-1} \mathbf{a})^{-1}$. For $N \rightarrow +\infty$, the true mixing vector is a stationary point only if (36) is fulfilled. The corresponding algorithm, similar to that proposed in [33] but leaving γ free, will be referred to as OGICE_a.

C. Preconditioning

The multiplicative form of the mixing model (3) allows us to consider the gradient computed according to transformed input signals $\mathbf{U} = \mathbf{D}\mathbf{X}$, where \mathbf{D} is a *preconditioning* non-singular matrix. We will consider the preconditioning applied within OGICE_w since this will help us reveal the connection between OGICE_w and the three well-known ICA/BSE algorithms.

Let $\mathbf{w}_{\mathbf{x}}$ and $\mathbf{w}_{\mathbf{u}}$ be the separating vectors operating on \mathbf{X} and \mathbf{U} , respectively, giving the same extracted signal, i.e., $\hat{\mathbf{s}} = \mathbf{w}_{\mathbf{x}}^H \mathbf{X} = \mathbf{w}_{\mathbf{u}}^H \mathbf{U}$. It follows that $\mathbf{w}_{\mathbf{x}} = \mathbf{D}^H \mathbf{w}_{\mathbf{u}}$. Consider now the gradient (33) when the input data is \mathbf{U} and the initial vector is $\mathbf{w}_{\mathbf{u}}$. Let that gradient be denoted by $\Delta_{\mathbf{u}}$. The sample covariance matrix of \mathbf{U} is $\hat{\mathbf{C}}_{\mathbf{u}} = \mathbf{D} \hat{\mathbf{C}}_{\mathbf{x}} \mathbf{D}^H$, so the right-hand side of (33) provides

$$\begin{aligned} \Delta_{\mathbf{u}} &= \frac{\hat{\mathbf{C}}_{\mathbf{u}} \mathbf{w}_{\mathbf{u}}}{\mathbf{w}_{\mathbf{u}}^H \hat{\mathbf{C}}_{\mathbf{u}} \mathbf{w}_{\mathbf{u}}} - \frac{1}{N} \mathbf{U} \phi(\hat{\mathbf{s}})^T \\ &= \frac{\mathbf{D} \hat{\mathbf{C}}_{\mathbf{x}} \mathbf{D}^H \mathbf{w}_{\mathbf{u}}}{\mathbf{w}_{\mathbf{u}}^H \mathbf{D} \hat{\mathbf{C}}_{\mathbf{x}} \mathbf{D}^H \mathbf{w}_{\mathbf{u}}} - \frac{1}{N} \mathbf{D} \mathbf{X} \phi(\hat{\mathbf{s}})^T \\ &= \mathbf{D} \left(\frac{\hat{\mathbf{C}}_{\mathbf{x}} \mathbf{w}_{\mathbf{x}}}{\mathbf{w}_{\mathbf{x}}^H \hat{\mathbf{C}}_{\mathbf{x}} \mathbf{w}_{\mathbf{x}}} - \frac{1}{N} \mathbf{X} \phi(\hat{\mathbf{s}})^T \right) = \mathbf{D} \Delta_{\mathbf{x}}, \end{aligned} \quad (38)$$

where $\Delta_{\mathbf{x}}$ denotes the ‘‘normal’’ gradient, that is, when the input data is \mathbf{X} and the initial vector is $\mathbf{w}_{\mathbf{x}}$. Note that (38) remains valid when the normalization of ϕ (dividing by ν) is taken into account, because ν is only a function of $\hat{\mathbf{s}}$.

After updating $\mathbf{w}_{\mathbf{u}}$ by $\mathbf{w}_{\mathbf{u}}^{\text{new}} = \mathbf{w}_{\mathbf{u}} + \mu \Delta_{\mathbf{u}}$, the extracted signal is equal to

$$\begin{aligned} (\mathbf{w}_{\mathbf{u}}^{\text{new}})^H \mathbf{U} &= (\mathbf{w}_{\mathbf{u}} + \mu \Delta_{\mathbf{u}})^H \mathbf{U} \\ &= (\mathbf{D}^{-H} \mathbf{w}_{\mathbf{x}} + \mu \mathbf{D} \Delta_{\mathbf{x}})^H \mathbf{D} \mathbf{X} \\ &= (\mathbf{w}_{\mathbf{x}} + \mu \mathbf{D}^H \mathbf{D} \Delta_{\mathbf{x}})^H \mathbf{X}. \end{aligned} \quad (39)$$

It follows that the gradient update computed on the preconditioned data \mathbf{U} corresponds to a modified update rule for $\mathbf{w}_{\mathbf{x}}$ given by

$$\mathbf{w}_{\mathbf{x}} \leftarrow \mathbf{w}_{\mathbf{x}} + \mu \mathbf{D}^H \mathbf{D} \Delta_{\mathbf{x}}. \quad (40)$$

For $\mathbf{D} = \mathbf{I}_d$, the modified update rule obviously coincides with the original one. In the following subsection, we will consider other special choices of \mathbf{D} and compare the modified OGICE_w with other ICA/BSE methods known in the literature.

D. Relation to Gradient and Natural Gradient ICA Methods

Here, OGICE_w is compared with the method by Bell and Sejnowski [52] for ICA and with its popular modification known as Amari’s Natural Gradient (NG) algorithm [53]; see also [54] and [51] for the complex-valued variant. In each step of Bell and Sejnowski’s method (BS), the whole de-mixing matrix is updated as

$$\Delta \mathbf{W} \leftarrow \mathbf{W}^{-H} - \overline{\phi(\mathbf{W}\mathbf{X})} \mathbf{X}^H / N. \quad (41)$$

After taking the conjugate transpose on both sides, and denoting $\mathbf{W}^{-1} = \mathbf{A}$, this update can be re-written as

$$\Delta \mathbf{W}^H \leftarrow \mathbf{A} - \mathbf{X}\phi(\mathbf{W}\mathbf{X})^T/N. \quad (42)$$

Now, the right-hand side of (33) corresponds to any row on the right-hand side of (42).

The de-mixing matrix update in NG is obtained when the right-hand side of (42) is multiplied by $\mathbf{W}^H \mathbf{W}$ from the left, which implies

$$\Delta \mathbf{W}^H \leftarrow \mathbf{W}^H (\mathbf{I}_d - \mathbf{W}\mathbf{X}\phi(\mathbf{W}\mathbf{X})^T/N). \quad (43)$$

OGICE_w becomes similar to NG when considering it with the modified update (40) where the precondition matrix is $\mathbf{D} = \mathbf{W}_{\text{ICE}}$. This choice corresponds to the update when the input data are pre-separated by the current de-mixing matrix prior to each iteration, and the starting point \mathbf{w} is equal to the unit vector (the first column of \mathbf{I}_d).

The main difference between OGICE_w and the respective ICA algorithms thus resides in the feature of BS and NG performing updates of the whole de-mixing matrix, while OGICE_w updates only its first row (the separating vector) under the orthogonal constraint. Next, the nonlinearity in OGICE_w is normalized according to (36), while neither BS nor NG makes use of any normalization.

E. Relation to One-Unit FastICA

One-unit FastICA (FICA) was derived as a fixed-point algorithm that minimizes the entropy of the extracted signal under the unit scale constraint. The FICA update for the separating vector can be written as [25]

$$\mathbf{w} \leftarrow \mathbf{w} - (\widehat{\mathbf{C}}_x^{-1} \mathbf{X}\psi(\widehat{\mathbf{s}})^T/N - \nu \mathbf{w})/(\rho - \nu), \quad (44)$$

$$\mathbf{w} \leftarrow \mathbf{w} / \sqrt{\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}}, \quad (45)$$

where $\nu = \widehat{\mathbf{s}}\psi(\widehat{\mathbf{s}})^T/N$ and $\rho = \psi'(\widehat{\mathbf{s}})\mathbf{1}_N/N$, where ψ' is the derivative of ψ , and $\mathbf{1}_N$ denotes the column vector consisting of ones with a length of N . Note that (45) corresponds to the normalization of \mathbf{w} so that the scale of the extracted signal equals one.

FICA is better known when it operates on pre-whitened data \mathbf{X} , which means that it is normalized prior to the optimization so that its sample covariance matrix is \mathbf{I}_d . This corresponds to the choice of the preconditioning matrix in Section III-C as $\mathbf{D} = \mathbf{F}\widehat{\mathbf{C}}_x^{-1/2}$, where $\widehat{\mathbf{C}}_x^{-1/2}$ denotes the inverse matrix square root of $\widehat{\mathbf{C}}_x$, and \mathbf{F} is an arbitrary unitary matrix. It then holds that $\mathbf{D}^H \mathbf{D} = \widehat{\mathbf{C}}_x^{-1}$, and we can compare the modified update rule of OGICE_w with (44). Specifically, the OGICE_w update modified according to (40) together with the nonlinearity normalization can be written as

$$\mathbf{w} \leftarrow \mathbf{w} + \mu \left(\frac{\mathbf{w}}{\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}} - \nu^{-1} \frac{1}{N} \widehat{\mathbf{C}}_x^{-1} \mathbf{X}\phi(\widehat{\mathbf{s}})^T \right) \quad (46)$$

where μ is the step length parameter. By comparing (44) and (46), the updates coincide when $\mu = \frac{\nu}{\rho - \nu}$, provided that $\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w} = 1$.

In conclusion, FICA and OGICE_w correspond to the same method when (a) the input data are pre-whitened (directly or through the preconditioning matrix and the modified update); (b) the step length in OGICE_w is selected adaptively as $\mu = \frac{\nu}{\rho - \nu}$; and (c) OGICE_w is forced to operate on the unit-scale sphere, which can be achieved through normalizing \mathbf{w} after each iteration as in (45). These results extend the analysis done in [55].

F. Switched Optimization

When all sources should be separated, as in ICA, it is less important which source is extracted in which output channel because all sources are separated in the end. However, when only one source should be extracted (based on the initial value of the mixing/separating vector), the size of the ROC becomes essential.

The ROC depends on the surface of the objective function for the given algorithm. This is influenced by all properties of the observed signals, namely, by the signals' distributions and their initial Signal-to-Interference Ratio (SIR) values. The SIR value of any signal that can be expressed as $y = \mathbf{q}^H \mathbf{u}$ is defined by

$$\text{SIR}_y = \frac{|q_1|^2 \mathbf{E}[|u_1|^2]}{\sum_{k=2}^d |q_k|^2 \mathbf{E}[|u_k|^2]} = \frac{|q_1|^2 \mathbf{E}[|s|^2]}{\sum_{k=2}^d |q_k|^2 \mathbf{E}[|u_k|^2]}, \quad (47)$$

where the expectation values are replaced by their sample-based averages when considering a finite number of samples. For example, the SIR value for the extracted signal corresponds to $\mathbf{q}^H = \mathbf{w}^H \mathbf{A}$; the SIR value in the k th input channel x_k corresponds to $\mathbf{q}^H = (\mathbf{A})_{k,:}$. It is seen that the initial SIR value on the k th channel is a function of the original signals' scales and of the mixing matrix. Its influence on the ROC is, however, difficult to analyze as it is different on each input channel.

Nevertheless, we can restrict our considerations to situations where the initial SIR value is approximately the same on all channels. This happens, for example, when the mutual distances of sensors are small compared to the distances of the sources from the sensor array. For now, let us assume that the magnitude of each element in the mixing matrix is approximately equal to a constant, so the initial SIR values are mainly influenced by the scales of the sources. It is then meaningful to introduce the so-called Scales Ratio (SR) related to the SOI by

$$\text{SR} = \frac{\mathbf{E}[|s|^2]}{\frac{1}{d-1} \sum_{i=2}^d \mathbf{E}[|u_i|^2]}. \quad (48)$$

The following example shows how the ROC of the OGICE algorithms can be influenced by SR.

Consider a situation where the SOI is a "weak" signal, i.e., $\text{SR} \ll 0$ dB. The mixing vector \mathbf{a} is then "hard" to find while the background subspace can be identified "easily". For the de-mixing matrix, the problem is reciprocal. The estimation of \mathbf{B} in (8) is inaccurate as \mathbf{B} purely depends on \mathbf{a} , while the estimation of \mathbf{w} yields a low variance.

Fig. 1 shows the objective function in the case of a real-valued mixture of two Laplacean components where one plays the role of the SOI and the other one is the background source (but the roles can be interchanged); the number of samples is

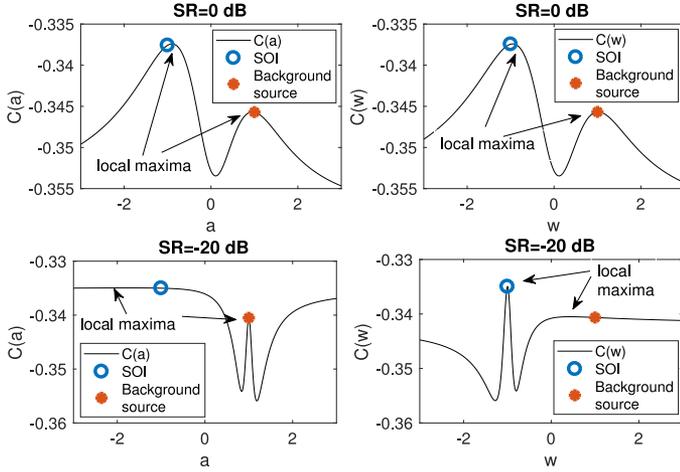


Fig. 1. Examples of the contrast function (26) as it depends on $\mathbf{a} = [1; a]$ and $\mathbf{w} = [1; w]$, respectively, when SR is 0 dB and -20 dB.

$N = 1000$; the mixing matrix is $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$; SR is considered in two settings: 0 dB and -20 dB, respectively. The function (26) is shown, as well as the way in which it depends on a and w , respectively, where the mixing vector is $\mathbf{a} = [1; a]$ and the separating vector is $\mathbf{w} = [1; w]$, respectively. The nonlinearity is $\log f(x) = -\log |\cosh(x/\sigma_x)|$ where σ_x^2 is the variance of the input. The perfect extraction of the SOI is achieved for $a = -1$ and $w = -1$, while $a = 1$ and $w = 1$ correspond to the extraction of the background source.

For SR = 0 dB, the surface of the contrast function as it depends on a or w , respectively, is almost the same. The local maxima are slightly distant from perfect solutions, and the sizes of ROC related to the maxima are approximately equal for both sources.

For SR = -20 dB, the background source dominates the mixture. Here, the maximum of $\mathcal{C}(a)$ corresponding to the mixing vector of the SOI is significantly distant from its ideal value, the function is almost flat in the vicinity of that maximum, and the corresponding ROC is wide. The separating vector for SOI is precisely localized by a sharp local maximum of $\mathcal{C}(w)$, which has a narrow ROC. The exact opposite is true for the maxima corresponding to the dominating background source.

In this example, OGICE_w is more advantageous when $\text{SR} \gg 0$ dB in the sense that the ROC corresponding to the SOI is wide. Even when the initialization of OGICE_w is significantly deviated, the probability of the successful convergence is high. Similarly, OGICE_a is advantageous when $\text{SR} \ll 0$ dB.

We have been working under the assumption that the elements of the mixing matrix have approximately the same magnitude. However, when this condition is not satisfied or once the input data are modified by a preconditioning matrix as considered in Section III-C, the initial SIR values can be different on different channels. Then, the analysis of influence of SR on ROC would be more difficult, although similar conclusions could be expected.

Nevertheless, the above observations suggest a practical algorithm for situations where the SIR value is approximately the same on all channels. It is based on an heuristic selection between the optimization in \mathbf{a} and \mathbf{w} .

Algorithm 2: OGICE_s: ICE Algorithm With Automatic Selection of Optimization Parameter.

Input: \mathbf{X} , \mathbf{a}_{ini} , μ , tol , Q

Output: \mathbf{a} , \mathbf{w}

```

1  $\hat{\mathbf{C}}_x = \mathbf{X}\mathbf{X}^H/N$ ;  $\mathbf{a} = \mathbf{a}_{\text{ini}}$ ;  $i = 0$ ;
2 repeat
3   if  $i \equiv 0 \pmod{Q}$  then
4      $\kappa = \mathcal{B}(\mathbf{a})$ ; /* using (50) */
5   end
6    $i \leftarrow i + 1$ ;
7   if  $\kappa < 0.1$  then
8     Update  $\mathbf{w}$  as in  $\text{OGICE}_w$ 
9      $\mathbf{a} \leftarrow (\mathbf{w}^H \hat{\mathbf{C}}_x \mathbf{w})^{-1} \hat{\mathbf{C}}_x \mathbf{w}$ ;
10  else
11    Update  $\mathbf{a}$  as in  $\text{OGICE}_a$ 
12     $\mathbf{w} \leftarrow (\mathbf{a}^H \hat{\mathbf{C}}_x^{-1} \mathbf{a})^{-1} \hat{\mathbf{C}}_x^{-1} \mathbf{a}$ ;
13  end
14 until  $\|\Delta\| < \text{tol}$ ;

```

Let \mathbf{a} be the current estimate of the mixing vector. Since the sequence of vectors $\hat{\mathbf{C}}_x \mathbf{a}$, $\hat{\mathbf{C}}_x^2 \mathbf{a}$, $\hat{\mathbf{C}}_x^3 \mathbf{a}, \dots$ (after each vector is normalized) quickly converges to the eigenvector of $\hat{\mathbf{C}}_x$ corresponding to the largest eigenvalue (for almost any initial value of \mathbf{a}),

$$\mathbf{b} = \hat{\mathbf{C}}_x \mathbf{a} \quad (49)$$

can be viewed as the first-order approximation of that eigenvector. The largest eigenvalue is therefore approximately equal to $\lambda_b = b_1/a_1$, and, consequently, $\|\mathbf{a}/a_1 - \mathbf{b}/\lambda_b\|$ is small when \mathbf{a} is close to the dominant eigenvector.

Next, to assess the dominance of that eigenvector, that is, whether it is significantly larger compared to the other eigenvectors, we propose to compute the ratio of norms of matrices $\hat{\mathbf{C}}_x - \lambda_b \mathbf{b}\mathbf{b}^H / \|\mathbf{b}\|^2$ and $\hat{\mathbf{C}}_x$. The ratio is small when \mathbf{b} is a dominant eigenvector of $\hat{\mathbf{C}}_x$. The criterion of “proximity and dominance” is therefore defined as

$$\mathcal{B}(\mathbf{a}) = \|\mathbf{a}/a_1 - \mathbf{b}/\lambda_b\| \frac{\|\hat{\mathbf{C}}_x - \lambda_b \mathbf{b}\mathbf{b}^H / \|\mathbf{b}\|^2\|_F}{\|\hat{\mathbf{C}}_x\|_F}. \quad (50)$$

The proposed algorithm, referred to as OGICE_s , selects the optimization parameter based on the current value of $\mathcal{B}(\mathbf{a})$. When $\mathcal{B}(\mathbf{a}) < \tau$, the optimization in \mathbf{w} is selected; otherwise, the optimization proceeds in \mathbf{a} . Normally, we select $\tau = 0.1$. To lower the computational load, the criterion is recomputed only once after Q iterations ($Q = 10$). The stopping condition of OGICE_s is the same as that in OGICE_w or OGICE_a ; see the summary in Algorithm 2.

IV. INDEPENDENT VECTOR EXTRACTION

A. Definition

In this section, we extend the ideas of the previous sections, which were derived for a single mixture, to joint extraction of an independent vector component from a set of K instantaneous mixtures.

Independent Vector Analysis (IVA) is defined as the goal to jointly separate K mixtures $\mathbf{x}^k = \mathbf{A}^k \mathbf{u}^k$, $k = 1, \dots, K$, where each mixture is defined as in (1). Compared to ICA, in IVA it is assumed that K -tuples of original sources, where each tuple contains one source from each mixture, are mutually independent but their elements can be dependent; the tuples are referred to as vector components [39].

In IVE, we re-parameterize each mixing matrix as in ICE in order to extract only one independent vector component. Let K instantaneous mixtures, each of the same dimension,⁴ be described by

$$\mathbf{x}^k = \mathbf{A}_{\text{ICE}}^k \mathbf{v}^k, \quad k = 1, \dots, K, \quad (51)$$

where $\mathbf{A}_{\text{ICE}}^k$ obeys the same structure as (3), and $\mathbf{v}^k = [s^k; \mathbf{z}^k]$. A joint mixture model can be written as

$$\mathbb{x} = \mathbb{A}_{\text{IVE}} \mathbb{v}. \quad (52)$$

The double-striking font will be used to denote concatenated variables or parameters from K mixtures, e.g., $\mathbb{x} = [\mathbf{x}^1; \dots; \mathbf{x}^K]$. The joint mixing matrix obeys $\mathbb{A}_{\text{IVE}} = \text{bdiag}(\mathbf{A}_{\text{ICE}}^1, \dots, \mathbf{A}_{\text{ICE}}^K)$, where $\text{bdiag}(\cdot)$ denotes a block-diagonal matrix with the arguments on the main block-diagonal.

Similar to (12), the joint statistical model is given by

$$p_{\mathbb{v}}(\mathbb{v}) = p_{\mathbb{s}}(\mathbb{s})p_{\mathbb{z}}(\mathbb{z}). \quad (53)$$

In this model, \mathbb{s} and \mathbb{z} are independent but the elements inside of them can be dependent. Like in the previous section, we constrain our considerations to the Gaussian background modeling; so it is assumed that

$$\mathbb{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_{\mathbb{z}}), \quad (54)$$

where $\mathbf{C}_{\mathbb{z}} = \mathbb{E}[\mathbb{z}\mathbb{z}^H]$ whose ij th block of dimension $(d-1) \times (d-1)$ is $\mathbf{C}_{\mathbb{z}}^{ij} = \mathbb{E}[\mathbf{z}^i \mathbf{z}^j H]$; similarly $\mathbf{C}_{\mathbb{x}}$ as well as the sample-based counterparts $\widehat{\mathbf{C}}_{\mathbb{z}}$ and $\widehat{\mathbf{C}}_{\mathbb{x}}$ are defined.

Similar to (26), the quasi-likelihood contrast function following from (53) (for one signal sample), is

$$\begin{aligned} \mathcal{J}(\mathbb{w}, \mathbb{a}) &= \log f(\widehat{\mathbf{s}}^1, \dots, \widehat{\mathbf{s}}^K) \\ &- \sum_{i=1}^K \sum_{j=1}^K \mathbf{x}^{iH} \mathbf{B}^j \mathbf{R}^{ij} \mathbf{B}^j \mathbf{x}^j + \sum_{k=1}^K \log |\det \mathbf{W}_{\text{ICE}}^k|^2, \end{aligned} \quad (55)$$

where $\widehat{\mathbf{s}}^k = (\mathbf{w}^k)^H \mathbf{x}^k$, and \mathbf{R}^{ij} is a weighting matrix substituting the ij th block of the unknown $\mathbf{C}_{\mathbb{z}}^{-1}$.

B. Gradient of the Contrast Function

After a lengthy computation, which follows steps similar to those described in Appendix C (we skip the details to save space), the derivative of \mathcal{J} with respect to $(\mathbf{w}^k)^H$ under the constraints (similar to (23))

$$\mathbf{a}^k = \frac{\widehat{\mathbf{C}}_{\mathbb{x}}^{kk} \mathbf{w}^k}{\mathbf{w}^{kH} \widehat{\mathbf{C}}_{\mathbb{x}}^{kk} \mathbf{w}^k}, \quad k = 1, \dots, K, \quad (56)$$

⁴IVE can be formulated such that each mixture involves a different number of original signals. However, we consider equal dimension for all K mixtures to simplify the exposition.

and when $\mathbf{R}^{k\ell}$ is selected as the $k\ell$ th block of $\widehat{\mathbf{C}}_{\mathbb{z}}^{-1}$, reads

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \mathbf{w}^{kH}} &= \mathbf{a}^k - \frac{1}{N} \mathbf{X}^k \phi^k(\widehat{\mathbf{s}}^1, \dots, \widehat{\mathbf{s}}^K)^T \\ &+ \frac{1}{\mathbf{w}^{kH} \widehat{\mathbf{C}}_{\mathbb{x}}^{kk} \mathbf{w}^k} \widehat{\mathbf{C}}_{\mathbb{x}}^{kk} \mathbf{B}^{kH} \boldsymbol{\epsilon}^k. \end{aligned} \quad (57)$$

Here, $\widehat{\mathbf{s}}^k = (\mathbf{w}^k)^H \mathbf{X}^k$,

$$\phi^k(\xi^1, \dots, \xi^K) = -\frac{\partial \log f(\xi^1, \dots, \xi^K)}{\partial \xi^k}, \quad (58)$$

is the score function related to the model joint density $f(\cdot)$ of \mathbb{s} with respect to the k th variable, and

$$\boldsymbol{\epsilon}^k = \sum_{\ell=1}^K \mathbf{R}^{k\ell} \boldsymbol{\theta}^{\ell k}, \quad \text{where } \boldsymbol{\theta}^{\ell k} = \widehat{\mathbf{Z}}^{\ell} (\widehat{\mathbf{s}}^k)^H / N. \quad (59)$$

By comparing (33) with (57), the latter differs only in that the nonlinearity $\phi^k(\cdot)$ is dependent on the SOIs separated from all K mixtures, plus the third term that does not occur in (33).

$\boldsymbol{\theta}^{\ell k}$ is the sample correlation between the estimated SOI in the k th mixture and the separated background in the ℓ th mixture, and can also be written as $\boldsymbol{\theta}^{\ell k} = \mathbf{B}^{\ell} \widehat{\mathbf{C}}_{\mathbb{x}}^{\ell k} \mathbf{w}^k$. For $k = 1, \dots, K$, $\boldsymbol{\theta}^{kk} = \mathbf{0}$ due to the OG, but, for $k \neq \ell$, $\boldsymbol{\theta}^{\ell k}$ is non-zero in general. Therefore, the third term in (57) vanishes only when $\mathbf{R}^{k\ell} = \mathbf{0}$ for $k \neq \ell$, $\ell = 1, \dots, K$.

Here, a special case is worth considering in which $\mathbf{C}_{\mathbb{x}} = \text{bdiag}(\mathbf{C}_{\mathbb{x}}^{11}, \dots, \mathbf{C}_{\mathbb{x}}^{KK})$; in other words, the mixtures (51) are uncorrelated (to each other), and there are only higher-order dependencies, if any. Then, $\mathbf{C}_{\mathbb{z}}$ has the same block-diagonal structure as $\mathbf{C}_{\mathbb{x}}$, so it is reasonable to select $\mathbf{R}^{k\ell} = \mathbf{0}$ for $k \neq \ell$, although the sample covariances $\widehat{\mathbf{C}}_{\mathbb{z}}^{k\ell}$ are not exactly zero. In that case, (57) is simplified to (compare with (33))

$$\frac{\partial \mathcal{J}}{\partial \mathbf{w}^{kH}} = \mathbf{a}^k - \frac{1}{N} \mathbf{X}^k \phi^k(\widehat{\mathbf{s}}^1, \dots, \widehat{\mathbf{s}}^K)^T. \quad (60)$$

This observation is in agreement with the literature. The joint separation of correlated mixtures can be achieved using second-order statistics only [6], [56]. Uncorrelated mixtures arise, for instance, in the frequency-domain separation of convolutive mixtures. Here, the non-Gaussianity and higher-order moments are necessary for separating the mixtures [9], [57].

C. Gradient Algorithms for IVE

The constrained gradient can, similarly to (60) and (37), be computed with respect to $(\mathbf{a}^k)^H$, which gives (we skip the detailed computation)

$$\frac{\partial \mathcal{J}}{\partial \mathbf{a}^{kH}} = \mathbf{w}^k - \frac{\lambda_{\mathbf{a}}^k}{N} (\widehat{\mathbf{C}}_{\mathbb{x}}^{kk})^{-1} \mathbf{X}^k \phi^k(\widehat{\mathbf{s}}^1, \dots, \widehat{\mathbf{s}}^K)^T, \quad (61)$$

where $\lambda_{\mathbf{a}}^k = (\mathbf{a}^{kH} (\widehat{\mathbf{C}}_{\mathbb{x}}^{kk})^{-1} \mathbf{a}^k)^{-1}$. Now, the gradient optimization algorithms for IVE (considering only sets of uncorrelated instantaneous mixtures) can proceed in the same way as those for ICE with the following differences:

- 1) In each iteration, \mathbf{w}_k or \mathbf{a}_k are updated by adding a step in the direction of the gradient (60) or (61), respectively, for each $k = 1, \dots, K$.

Algorithm 3: OGIVE_w: Orthogonally Constrained Extraction of an Independent Vector Component From the Set of Mutually Uncorrelated Mixtures $\mathbf{X}^1, \dots, \mathbf{X}^K$.

Input: $\mathbf{X}^k, \mathbf{w}_{\text{ini}}^k, k = 1, \dots, K, \mu, \text{tol}$
Output: $\mathbf{a}^k, \mathbf{w}^k, k = 1, \dots, K$

```

1 foreach  $k = 1, \dots, K$  do
2    $\hat{\mathbf{C}}_{\mathbf{x}}^{kk} = \mathbf{X}^k (\mathbf{X}^k)^H / N;$ 
3    $\mathbf{w}^k = \mathbf{w}_{\text{ini}}^k;$ 
4 end
5 repeat
6   foreach  $k = 1, \dots, K$  do
7      $\mathbf{a}^k \leftarrow ((\mathbf{w}^k)^H \hat{\mathbf{C}}_{\mathbf{x}}^{kk} \mathbf{w}^k)^{-1} \hat{\mathbf{C}}_{\mathbf{x}}^{kk} \mathbf{w}^k;$ 
8      $\hat{\mathbf{s}}^k \leftarrow (\mathbf{w}^k)^H \mathbf{X}^k;$ 
9   end
10  foreach  $k = 1, \dots, K$  do
11     $\nu^k \leftarrow \hat{\mathbf{s}}^k \phi^k(\hat{\mathbf{s}}^1, \dots, \hat{\mathbf{s}}^K)^T / N;$ 
12     $\Delta^k \leftarrow \mathbf{a}^k - (\nu^k)^{-1} \mathbf{X}^k \phi^k(\hat{\mathbf{s}}^1, \dots, \hat{\mathbf{s}}^K)^T / N;$ 
13     $\mathbf{w}^k \leftarrow \mathbf{w}^k + \mu \Delta^k;$ 
14  end
15 until  $\max\{\|\Delta^1\|, \dots, \|\Delta^K\|\} < \text{tol};$ 

```

- 2) The nonlinear functions $\phi^k, k = 1, \dots, K$, depend on the current outputs of all K iterative algorithms, which fact makes them mutually dependent.

We will refer to these algorithms as to OGIVE_w and OGIVE_a, respectively; for illustration, OGIVE_w is described in Algorithm 3.

Finally, we introduce a method referred to as OGIVE_s, where the idea presented in Section III-F is applied within each mixture. The parameter in which the gradient optimization proceeds is selected based on the heuristic criterion (50). It is important to note that the selection can be different for each mixture.

In fact, this approach inherently assumes that the behavior of the optimization process within any mixture is similar to the case of ICE. The behavior can be different in the case of IVE as the parallel extraction algorithms influence each other. The advantage of this feature is that the dependence can bring a synergic effect: When most initial separating/mixing vectors lie in the ROC of the SOI, the convergence within the other mixtures can be enforced.

V. SIMULATIONS

In simulations,⁵ we focus on the sensitivity of the proposed ICE and IVE algorithms with respect to the initialization and compare them with other methods. In one simulated trial, K instantaneous mixtures are generated with random mixing matrices $\mathbf{A}^k, k = 1, \dots, K$. The SOIs for these mixtures are obtained as K signals drawn independently from the Laplacean distribution and mixed by a random unitary matrix; hence, they are uncorrelated and dependent. The background is obtained by generating independent components u_2^k, \dots, u_d^k from the Gaus-

sian or Laplacean distribution, $k = 1, \dots, K$; all the distributions are circular.

For each mixture, the SR is selected randomly, either -10 or 10 dB. The mixing matrices are drawn from the uniform distribution; the real part in $[1, 2]$ and the imaginary part in $[0, 1]$. This choice helps us keep the initial SIR values approximately equal across all input channels, that is, less dependent on the mixing matrix while mostly dependent on the SR.

The comparison involves One-Unit FastICA (FICA) [59], three variants of OGICE proposed in Section III, the Natural Gradient algorithm (NG) [53] and its scaled version (scNG) [60], which is frequently used in audio separation methods, as well as three variants of OGIVE proposed in Section IV-C. These algorithms are initialized by $\mathbf{a}_{\text{ini}} = \mathbf{a} + \mathbf{e}_{\text{ini}}$, where \mathbf{a} is the true mixing vector, and \mathbf{e}_{ini} is a random vector which is orthogonal to \mathbf{a} , and $\|\mathbf{e}_{\text{ini}}\|^2 = \epsilon^2$. The algorithms NG and scNG are initialized by the de-mixing matrix whose first row is equal to (22) where $\mathbf{a} = \mathbf{a}_{\text{ini}}$; the other rows are selected as in (8), which means that the initial background subspace is orthogonal to the initial SOI estimate.

Next, we also evaluate the SOI estimates obtained through (22) for \mathbf{a} equal to the true mixing vector (MPDR_oracle) and for $\mathbf{a} = \mathbf{a}_{\text{ini}}$ (MPDR_ini). While the performance of the MPDR_oracle gives an upper bound, the performance of MPDR_ini corresponds to a “do-nothing” solution purely relying on the initialization.

ICE/ICA methods are applied to each mixture separately, while IVE algorithms treat all K mixtures jointly.⁶ The nonlinearity $\phi(\xi) = \overline{\tanh}(\xi)$ is selected in the variants of OGICE and NG. FICA is used with $(1 + |\xi|^2)^{-1}$. For IVE algorithms, the choice is

$$\phi^k(\xi^1, \dots, \xi^K) = \overline{\tanh}(\xi^k) / \sqrt{\sum_{\ell=1}^K |\xi^\ell|^2}. \quad (62)$$

The problem of choosing an appropriate nonlinearity for the given method would go beyond the scope of this paper; see, e.g., [28], [49].

For the sake of completeness, we also include a semi-blind variant of OGIVE_s, which is modified in a way similar to that proposed in [61]. Specifically, a “pilot” component p is assumed to be available such that the SOIs within the K mixtures are dependent on it (usually there are only higher-order dependencies; see [9]). OGIVE_s is modified only by adding the $K + 1$ th variable into (62), which is $\xi_{K+1} = p$. In simulations, p is a random mixture of the SOIs. This method will be referred to as “Piloted OGIVE_s”.

The detailed settings of the compared algorithms are shown in Table I; these values were selected to ensure good performance of the methods.

A. Results

The algorithms were tested in 1000 independent trials for $d = 6, K = 4$, and $N = 1000$. Each extracted signal was assessed

⁵Matlab codes and results of the experiment are available at [58].

⁶IVE can take advantage of the dependence among SOIs from different mixtures while ICA/ICE cannot. One goal of our experiment here is to evaluate the improvement due to the joint extraction.

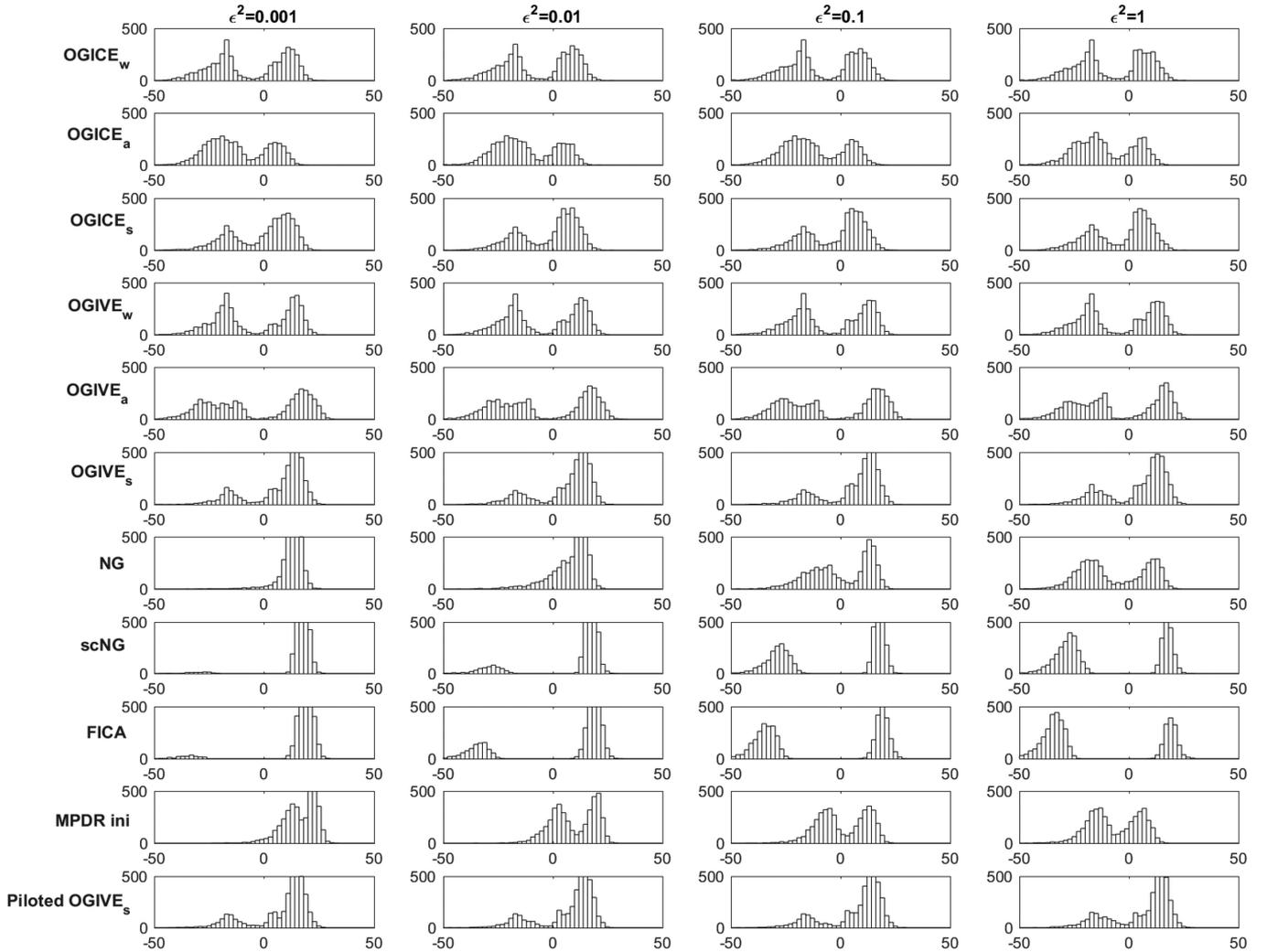


Fig. 2. Histograms of the output SIR values (in dB) achieved by the blind algorithms in 1000 trials (4000 extractions) when the background is Laplacean.

TABLE I
DETAILED SETTINGS OF THE COMPARED METHODS

Algorithm(s)	maximum # iterations	stopping threshold	step length
OGICE variants	5×10^3	10^{-3}	0.1
OGIVE variants	4×10^3	10^{-3}	0.1
NG, scNG	5×10^3	10^{-3}	0.02
FICA	1×10^3	10^{-6}	-

by the output SIR value (the ratio of powers of the SOI and of the other signals within the extracted signal). In the experiment, the achieved SIR values range from -50 to 50 dB depending on whether the SOI was extracted or a different source, and the SIR value also depends on the qualities of the given algorithm. Complete results achieved by the methods, for the Laplacean background, are shown as histograms in Fig. 2.

Since our primary focus is on the extraction of the SOI, the results in Fig. 3 show the percentages of successful extractions of the SOI (success rates) as functions of the initial error $\|e_{ini}\|^2 = \epsilon^2$. Here, each extraction is classed as *successful* if the output SIR value is greater than 0 dB. We first discuss these results as follows.

- 1) *MPDR*: *MPDR* oracle achieves a 100% success rate. The success rate of *MPDR_ini* is decaying with growing ϵ^2 as *MPDR_ini* yields the SIR value corresponding to the extracted signal using the initial mixing vector. Its success rate approaches 50% as $\epsilon^2 \rightarrow 1$, which corresponds to the fact that $SR = 10$ dB in about 50% of trials. The results of *MPDR_oracle* and of *MPDR_ini* do not depend on the signals' distributions.
- 2) *OGICE_{a/w}*: *OGICE_a* and *OGICE_w* achieve a success rate of 20–50% almost independently of ϵ^2 . This corresponds to the fact that maximum 50% of trials are advantageous for each method ($SR = -10$ dB for *OGICE_a* and $SR = +10$ dB for *OGICE_a*). The success rates are mostly lower than 50%, which shows that the mixing matrix also has an influence on the ROC of the SOI, so, for example, $SR = +10$ dB does not always guarantee that *OGICE_w* converges to the SOI from almost any initial value.⁷ For

⁷It can be verified, by repeating this experiment, that with $SR = \pm 15$ dB, *OGICE_{a/w}* achieve a success rate of almost 50%. It means that, with a higher range of SR, the influence of the mixing matrix (generated in the same way as in this experiment) is lower.

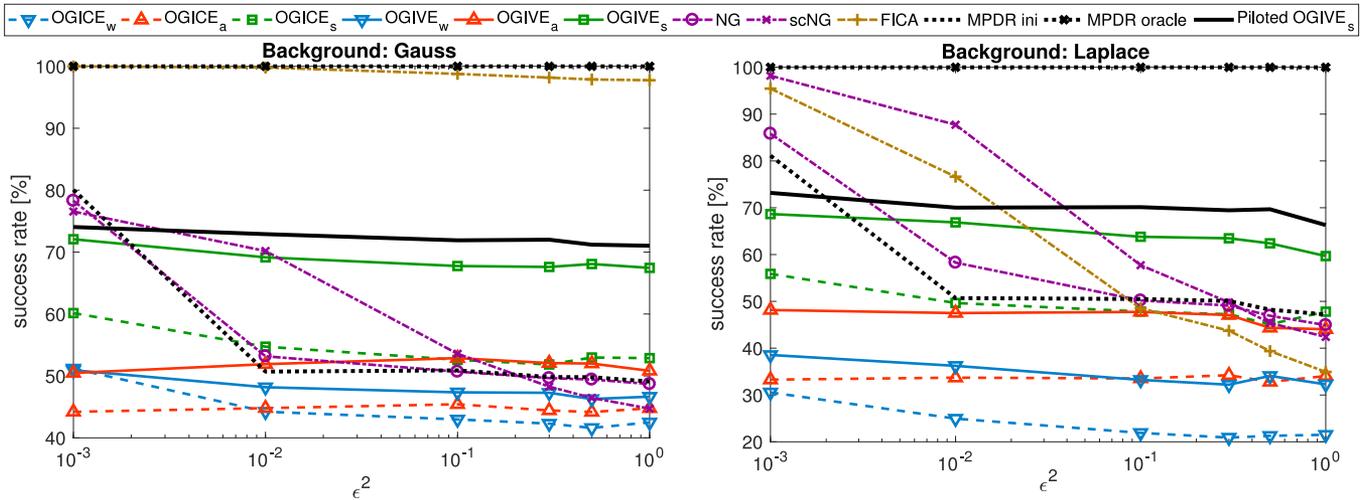


Fig. 3. Success rates as functions of the initial mean squared error, ϵ^2 for $d = 6$, $K = 4$, and $N = 1000$. The experiment was repeated in 1000 trials ($1000 \times K$ mixtures). The SOI is generated from Laplace components while the background is Gaussian or Laplacean.

the Laplacean background, the success rate values are lower than for the Gaussian background because there is a higher probability that the methods are attracted by a different non-Gaussian source.

- 3) $OGIVE_{a/w}$: The IVE counterparts of $OGICE_a$ and $OGICE_w$ perform similarly but their success rate values are always closer to 50%. This is caused by the joint extraction. $OGIVE_a$ and $OGIVE_w$ take advantage of the dependence between the SOIs in K mixtures, which improves the overall convergence.
- 4) *Switched optimization*: In terms of the success rate, $OGICE_s$ and $OGIVE_s$ achieve significant improvements compared to $OGICE_{a/w}$ and $OGIVE_{a/w}$, respectively. This indicates effectiveness of the decision rule based on the criterion (50) and to the synergy of the joint separation in the case of $OGIVE_s$. It is worth pointing out that the performances of $OGICE_s$ and $OGIVE_s$ slightly decrease with growing ϵ^2 . This is because for higher ϵ^2 , the initial mixing vector is more distant from the true value, so the criterion (50) is less reliable for selecting the optimization parameter. To avoid this drawback, a better decision rule is needed.
- 5) *NG and scNG*: The success rate of these ICA algorithms significantly depends on the initial error as well as on the distribution of the background. The success rate is superior for very small values of ϵ^2 , but it rapidly drops with growing ϵ^2 (scNG appears to be less sensitive than NG). The results also show that very good convergence is achieved for $\epsilon^2 < 10^{-2}$ when the background is Laplacean. This indicates that the ICA algorithms can take advantage of the non-Gaussianity of background signals.
- 6) *FICA* shows excellent results when the background is Gaussian. Since it is a fixed-point algorithm, it has a good ability to avoid shallow extremes of the contrast function that correspond to Gaussian components. Therefore, its global convergence is very good when there are no non-Gaussian components other than the SOI. For the same

reason, the success rate of FICA goes down with growing ϵ^2 when the background is Laplacean.

- 7) *Piloted $OGIVE_s$* : This algorithm gives a higher success rate than $OGIVE_s$ as it exploits the pilot-dependent component to keep converging to the SOI. Its performance is slightly decreasing when ϵ^2 grows due to the shortage of the criterion (50), as mentioned in Item 4 above.

Now we go back to the histograms of the output SIR values (in dB) shown in Fig. 2. Typically, these values are concentrated around two peaks with a central value of $\approx \pm 20$ dB. When more values are concentrated around a positive SIR value for any ϵ^2 , the given method shows good robustness against the initialization error as it mostly tends to keep converging to the SOI. From this perspective, $OGICE_s$, $OGIVE_s$ and *Piloted $OGIVE_s$* show the best results, as was already discussed above.

The variance of the SIR values around the peaks reflect the ability of the method to avoid local extremes of the contrast function and/or its ability to converge before the maximum number of iterations is reached. In this respect, scNG and FICA achieve superior results as they rarely yield an output SIR value within the range of $[-10, 10]$ dB. It is worth pointing out that scNG and FICA could be interpreted as gradient methods using special preconditioning and adaptive step lengths, as discussed in Sections III-C-III-E. The other compared algorithms, the variants of $OGICE$ and of $OGIVE$, utilize constant step lengths and do not apply any preconditioning. Their histograms are less concentrated around the main peaks, which means that they sometimes get stuck in a local extreme or converge too slowly to achieve the desired extreme of the contrast function.

VI. CONCLUSIONS

We have revised the problem of blind source extraction of an independent target source from background signals. The maximum likelihood approach where the mixing model is parameterized for the extraction of one source and where the background signals are modeled as a Gaussian mixture has been introduced

as Independent Component Extraction (ICE). Similarly, Independent Vector Extraction (IVE) has been introduced for the joint extraction problem.

Several variants of gradient algorithms have been derived. Our attention has been focused on their region of convergence related to the source of interest. It was shown that the size of the ROC is not only algorithm-dependent but it also strongly depends on the ratio of scales of the sources within the mixture. In particular, we have shown that the size of the ROC can be influenced through the selection of optimization parameters. This was corroborated by simulations where the methods endowed with the automatic selection of optimization parameters achieved a high rate of successful extractions of the SOI, almost independent of the initialization. Matlab codes of algorithms proposed in this article are available at [58].

The simulation study has also confirmed that the joint extraction through IVE brings advantageous features compared to ICE. In particular, the IVE methods with automatic selection show synergistic convergence, by which we mean that simultaneous convergence of the SOIs within several mixtures helps us enforce the convergence in the other mixtures as well.

Next, the gradient algorithms derived based on ICE have been compared with the Natural Gradient-based methods for ICA and with One-Unit FastICA. Close relationships between these methods have been revealed, which sheds light on the relation between ICE and the well-known blind source separation/extraction (BSS/BSE) methods. In particular, the importance of the orthogonal constraint (the orthogonality of the subspaces spanned by the SOI and the background signals) and the Gaussian modeling of the background signals in ICE methods has been shown. Therefore, future work should be focused on these aspects in order to improve overall properties of ICE/IVE algorithms in non-Gaussian backgrounds and in underdetermined scenarios.

APPENDIX A PROOF OF (22) AND (23)

Since $\mathbf{y} = \mathbf{Qz} = \mathbf{QBx}$, we can introduce the projection operator $\mathbf{\Pi}_y = \mathbf{QB}$, which is equal to

$$\mathbf{\Pi}_y = \mathbf{I}_d - \mathbf{a}\mathbf{w}^H. \quad (63)$$

According to (21), the OC can be written as

$$\widehat{\mathbf{Z}}^H \widehat{\mathbf{s}}/N = \widehat{\mathbf{B}}\widehat{\mathbf{C}}_x \mathbf{w} = \mathbf{0}. \quad (64)$$

By multiplying the latter equation from the left by \mathbf{Q} and using (63), we arrive at

$$(\mathbf{I}_d - \mathbf{a}\mathbf{w}^H)\widehat{\mathbf{C}}_x \mathbf{w} = \mathbf{0}, \quad (65)$$

$$\mathbf{w} - \widehat{\mathbf{C}}_x^{-1} \mathbf{a} (\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}) = \mathbf{0}. \quad (66)$$

Multiplying (66) from the left by \mathbf{a}^H gives

$$\mathbf{a}^H \mathbf{w} - \mathbf{a}^H \widehat{\mathbf{C}}_x^{-1} \mathbf{a} \mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w} = \mathbf{0}, \quad (67)$$

and since $\mathbf{a}^H \mathbf{w} = 1$, it holds that

$$\mathbf{a}^H \widehat{\mathbf{C}}_x^{-1} \mathbf{a} = (\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w})^{-1}. \quad (68)$$

By putting (66) and (68) together, (22) and (23) follow. ■

APPENDIX B PROOF OF (25)

Assume that \mathbf{a} is equal to its true value in (2) (hence $\mathbf{A} = \mathbf{A}_{\text{ICE}}$ and $\mathbf{W} = \mathbf{W}_{\text{ICE}}$), and recall that, in the determined ICE model, $\mathbf{y} = \mathbf{Qz}$ holds. Then, $\mathbf{w}_{\text{MPDR}}^H \mathbf{x} = s + \mathbf{w}_{\text{MPDR}}^H \mathbf{y}$. We should show that $\mathbf{a}^H \mathbf{C}_x^{-1} \mathbf{y} = \mathbf{0}$. It holds that

$$\mathbf{a}^H \mathbf{C}_x^{-1} \mathbf{y} = \mathbf{a}^H \mathbf{C}_x^{-1} \mathbf{Qz} \quad (69)$$

$$= \mathbf{a}^H (\mathbf{A}\mathbf{C}_v \mathbf{A}^H)^{-1} \mathbf{Qz} \quad (70)$$

$$= \mathbf{a}^H \mathbf{W}^H \mathbf{C}_v^{-1} \mathbf{W}\mathbf{Qz}, \quad (71)$$

where $\mathbf{C}_v = \mathbb{E}[\mathbf{v}\mathbf{v}^H]$. Next, it holds that $\mathbf{a}^H \mathbf{W}^H = [1, \mathbf{0}^H]$, and $\mathbf{W}\mathbf{Q} = [\mathbf{0}, \mathbf{I}_{d-1}]^H$. By taking into account the block-diagonal structure of \mathbf{C}_v , i.e.,

$$\mathbf{C}_v = \begin{pmatrix} \sigma_s^2 & \mathbf{0}^H \\ \mathbf{0} & \mathbf{C}_z \end{pmatrix}, \quad (72)$$

where σ_s^2 denotes the variance of s , (25) follows. ■

APPENDIX C COMPUTATION OF (27)

The following identities hold under the constraint (23).

$$\mathbf{g} = \mathbf{E} \mathbf{a} = \frac{\mathbf{E}\widehat{\mathbf{C}}_x \mathbf{w}}{\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}}, \quad (73)$$

$$\gamma = \mathbf{e}_1^H \mathbf{a} = \frac{\mathbf{e}_1^H \widehat{\mathbf{C}}_x \mathbf{w}}{\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}}, \quad (74)$$

$$1 - \mathbf{h}^H \mathbf{g} = \bar{\beta} \gamma = \bar{\beta} \frac{\mathbf{e}_1^H \widehat{\mathbf{C}}_x \mathbf{w}}{\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}}. \quad (75)$$

To derive (27), we proceed by computing the derivatives of the three terms in (26). First, using (28), it follows that

$$\frac{\partial}{\partial \mathbf{w}^H} \log f(\mathbf{w}^H \mathbf{x}) = -\phi(\mathbf{w}^H \mathbf{x})\mathbf{x}, \quad (76)$$

so that

$$\frac{1}{N} \sum_{n=1}^N -\phi(\mathbf{w}^H \mathbf{x}(n))\mathbf{x}(n) = -\frac{1}{N} \mathbf{X}\phi(\mathbf{w}^H \mathbf{X})^T, \quad (77)$$

where $\phi(\cdot)$ is applied element-wise in case of the vector argument.

Let \mathbf{x} be partitioned as $\mathbf{x} = [x_1; \mathbf{x}_2]$. Under the constraint (23) and using $\mathbf{B} = [\mathbf{g}, -\gamma\mathbf{I}_{d-1}]$, the second term in (26) (where the argument n is omitted) can be re-written as

$$\begin{aligned} \mathbf{x}^H \mathbf{B}^H \mathbf{R}\mathbf{B}\mathbf{x} &= (\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w})^{-1} \\ &\times (|x_1|^2 \mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{E}^H \mathbf{R}\mathbf{E} \widehat{\mathbf{C}}_x \mathbf{w} - \overline{x_1} \mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{E}^H \mathbf{e}_1^H \widehat{\mathbf{C}}_x \mathbf{w} \mathbf{R}\mathbf{x}_2 \\ &- x_1 \mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{e}_1 \mathbf{x}_2^H \mathbf{R}\mathbf{E} \widehat{\mathbf{C}}_x \mathbf{w} + \mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{e}_1 \mathbf{x}_2^H \mathbf{R}\mathbf{e}_1^H \widehat{\mathbf{C}}_x \mathbf{w} \mathbf{x}_2). \end{aligned} \quad (78)$$

By taking the derivative under the OG and after some rearrangements,

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}^H} \mathbf{x}^H \mathbf{B}^H \mathbf{R}\mathbf{B}\mathbf{x} &= -2\mathbf{a} \mathbf{x}^H \mathbf{B}^H \mathbf{R}\mathbf{B}\mathbf{x} \\ &+ (\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w})^{-1} \times (\overline{x_1} \widehat{\mathbf{C}}_x \mathbf{E}^H \mathbf{R}\mathbf{B}\mathbf{x} - \mathbf{x}_2^H \mathbf{R}\mathbf{B}\mathbf{x} \widehat{\mathbf{C}}_x \mathbf{e}_1). \end{aligned} \quad (79)$$

By considering the averages of the above terms over N samples, we arrive at the following chain of identities:

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^N \mathbf{x}(n)^H \mathbf{B}^H \mathbf{R} \mathbf{B} \mathbf{x}(n) &= \frac{1}{N} \sum_{n=1}^N \text{tr}(\mathbf{B}^H \mathbf{R} \mathbf{B} \mathbf{x}(n) \mathbf{x}(n)^H) \\ &= \text{tr}(\mathbf{R} \mathbf{B} \widehat{\mathbf{C}}_x \mathbf{B}^H) = \text{tr}(\mathbf{R} \widehat{\mathbf{C}}_z). \end{aligned} \quad (80)$$

Next,

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^N \overline{x_1}(n) \widehat{\mathbf{C}}_x \mathbf{E}^H \mathbf{R} \mathbf{B} \mathbf{x}(n) &= \widehat{\mathbf{C}}_x \mathbf{E}^H \mathbf{R} \mathbf{B} \mathbf{X} \mathbf{X}^H \mathbf{e}_1 / N \\ &= \widehat{\mathbf{C}}_x \mathbf{E}^H \widehat{\mathbf{R}} \widehat{\mathbf{Z}} \mathbf{X}^H \mathbf{e}_1 / N \\ &= \widehat{\mathbf{C}}_x \mathbf{E}^H \widehat{\mathbf{R}} \widehat{\mathbf{Z}} \left(\widehat{\mathbf{s}}^H \widehat{\mathbf{Z}}^H \right) \mathbf{A}_{\text{ICE}}^H \mathbf{e}_1 / N \\ &= \widehat{\mathbf{C}}_x \mathbf{E}^H \mathbf{R} \left(\mathbf{0}^H \widehat{\mathbf{C}}_z \right) \mathbf{A}_{\text{ICE}}^H \mathbf{e}_1 \\ &= \widehat{\mathbf{C}}_x \mathbf{E}^H \mathbf{R} \widehat{\mathbf{C}}_z \mathbf{h}, \end{aligned} \quad (81)$$

where we used (20) and (21). The last identity is

$$\frac{1}{N} \sum_{n=1}^N \mathbf{x}_2^H \mathbf{R} \mathbf{B} \mathbf{x} \widehat{\mathbf{C}}_x \mathbf{e}_1 = \text{tr}(\mathbf{R} \mathbf{B} \widehat{\mathbf{C}}_x \mathbf{E}^H) \widehat{\mathbf{C}}_x \mathbf{e}_1. \quad (82)$$

The derivative of the third term in (26) reads

$$\begin{aligned} (d-2) \frac{\partial}{\partial \mathbf{w}^H} \log |\gamma|^2 &= (d-2) \frac{\partial}{\partial \mathbf{w}^H} (\log |\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{e}_1|^2 - \log |\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}|^2) \\ &= (d-2) \frac{\partial}{\partial \mathbf{w}^H} (\log \mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{e}_1 - 2 \log \mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}) \\ &= (d-2) \left(\frac{\widehat{\mathbf{C}}_x \mathbf{e}_1}{\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{e}_1} - 2 \frac{\widehat{\mathbf{C}}_x \mathbf{w}}{\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}} \right) \\ &= (d-2) \left(\overline{\gamma}^{-1} \frac{\widehat{\mathbf{C}}_x \mathbf{e}_1}{\mathbf{w}^H \widehat{\mathbf{C}}_x \mathbf{w}} - 2\mathbf{a} \right). \end{aligned} \quad (83)$$

Now, (27) is obtained by putting together (77), (79), and (83) using the identities (80), (81), and (82). ■

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their very insightful and detailed comments that significantly helped to improve this paper.

REFERENCES

- [1] T.-W. Lee, *Independent Component Analysis—Theory and Applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1998.
- [2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Hoboken, NJ, USA: Wiley, 2001.
- [3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. Hoboken, NJ, USA: Wiley, 2002.
- [4] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications* (ser. Independent Component Analysis and Applications Series). Amsterdam, The Netherlands: Elsevier Science, 2010.
- [5] P. Smaragdakis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21–34, 1998.
- [6] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun, “Joint blind source separation by multiset canonical correlation analysis,” *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3918–3929, Oct. 2009.
- [7] D. Lahat and C. Jutten, “Joint blind source separation of multidimensional components: Model and algorithm,” in *Proc. 22nd Eur. Signal Process. Conf.*, Sep. 2014, pp. 1417–1421.
- [8] X.-L. Li, T. Adali, and M. Anderson, “Joint blind source separation by generalized joint diagonalization of cumulant matrices,” *Signal Process.*, vol. 91, no. 10, pp. 2314–2322, 2011.
- [9] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 70–79, Jan. 2007.
- [10] H. Sawada, R. Mukai, S. Araki, and S. Makino, “Solving the permutation and the circularity problem of frequency-domain blind source separation,” in *Proc. 18th Int. Congr. Acoust.*, Apr. 2004, pp. 1–89–1–92.
- [11] Z. Koldovský, J. Málek, P. Tichavský, and F. Nesta, “Semi-blind noise extraction using partially known position of the target source,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2029–2041, Oct. 2013.
- [12] H. Sawada, S. Araki, R. Mukai, and S. Makino, “Blind extraction of a dominant source signal from mixtures of many sources,” in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, vol. III, Mar. 2005, pp. 61–64.
- [13] S. Javidi, D. P. Mandic, and A. Cichocki, “Complex blind source extraction from noisy mixtures using second-order statistics,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 7, pp. 1404–1416, Jul. 2010.
- [14] S. Amari and A. Cichocki, “Adaptive blind signal processing-neural network approaches,” *Proc. IEEE*, vol. 86, no. 10, pp. 2026–2048, Oct. 1998.
- [15] S. A. Cruces-Alvarez, A. Cichocki, and S. Amari, “From blind signal extraction to blind instantaneous signal separation: Criteria, algorithms, and stability,” *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 859–873, Jul. 2004.
- [16] P. J. Huber, “Projection pursuit,” *Ann. Statist.*, vol. 13, no. 2, pp. 435–475, Jun. 1985.
- [17] M. Girolami and C. Fyfe, “Extraction of independent signal sources using a deflationary exploratory projection pursuit network with lateral inhibition,” *IEE Proc.—Vision, Image Signal Process.*, vol. 144, no. 5, pp. 299–306, Oct. 1997.
- [18] O. Shalvi and E. Weinstein, “Super-exponential methods for blind deconvolution,” *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 504–519, Mar. 1993.
- [19] J. J. Shynk and R. P. Gooch, “The constant modulus array for cochannel signal copy and direction finding,” *IEEE Trans. Signal Process.*, vol. 44, no. 3, pp. 652–660, Mar. 1996.
- [20] E. Moreau and O. Macchi, “High-order contrasts for self-adaptive source separation,” *Int. J. Adaptive Control Signal Process.*, vol. 10, no. 1, pp. 19–46, 1996.
- [21] T. Cover and J. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.
- [22] J.-F. Cardoso, “Blind signal separation: statistical principles,” *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.
- [23] N. Delfosse and P. Loubaton, “Adaptive blind separation of independent sources: A deflation approach,” *Signal Process.*, vol. 45, no. 1, pp. 59–83, 1995.
- [24] A. T. Erdogan, “On the convergence of ICA algorithms with symmetric orthogonalization,” *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2209–2221, Jun. 2009.
- [25] A. Hyvärinen and E. Oja, “A fast fixed-point algorithm for independent component analysis,” *Neural Comput.*, vol. 9, no. 7, pp. 1483–1492, Jul. 1997.
- [26] A. Hyvärinen, “Fast and robust fixed-point algorithm for independent component analysis,” *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.
- [27] T. Wei, “A convergence and asymptotic analysis of the generalized symmetric FastICA algorithm,” *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6445–6458, Dec. 2015.
- [28] 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.
- [29] A. Hyvärinen, “One-unit contrast functions for independent component analysis: A statistical analysis,” in *Proc. IEEE Signal Process. Soc. Workshop Neural Netw. Signal Process. VII*, Sep. 1997, pp. 388–397.
- [30] J.-F. Cardoso, “On the performance of orthogonal source separation algorithms,” in *Proc. Eur. Signal Process. Conf.*, Sep. 1994, pp. 776–779.

- [31] Z. Koldovský, P. Tichavský, and E. Oja, "Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér–Rao lower bound," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1265–1277, Sep. 2006.
- [32] D.-T. A. Pham, "Blind partial separation of instantaneous mixtures of sources," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation*. Berlin, Germany: Springer, 2006, pp. 868–875.
- [33] Z. Koldovský, P. Tichavský, and V. Kautský, "Orthogonally constrained independent component extraction: Blind MPDR beamforming," in *Proc. Eur. Signal Process. Conf.*, Sep. 2017, pp. 1195–1199.
- [34] J.-F. Cardoso, "On extracting the cosmic microwave background from multi-channel measurements," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Feb. 2017, pp. 403–413.
- [35] J. Dick, M. Remazeilles, and J. Delabrouille, "Impact of calibration errors on CMB component separation using FastICA and ILC," *Monthly Notices Roy. Astron. Soc.*, vol. 401, no. 3, pp. 1602–1612, 2010.
- [36] M. Remazeilles, J. Delabrouille, and J.-F. Cardoso, "Foreground component separation with generalized internal linear combination," *Monthly Notices Roy. Astron. Soc.*, vol. 418, no. 1, pp. 467–476, 2011.
- [37] 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.
- [38] V. Kautský, Z. Koldovský, and P. Tichavský, "Cramér-Rao-induced bound for interference-to-signal ratio achievable through non-Gaussian independent component extraction," in *Proc. IEEE 7th Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, Dec. 2017, pp. 1–4.
- [39] I. Lee, T. Kim, and T.-W. Lee, "Independent vector analysis for convolutive blind speech separation," in *Blind Speech Separation*. New York, NY, USA: Springer, 2007, pp. 169–192.
- [40] T. Adali, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 18–33, May 2014.
- [41] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Hoboken, NJ, USA: Wiley, 2002.
- [42] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, vol. 4, May 1998, pp. 1941–1944.
- [43] A. Hyvärinen, and U. Köster, "FastISA: A fast fixed-point algorithm for independent subspace analysis," in *Proc. 14th Eur. Symp. Artif. Neural Netw.*, 2006, pp. 371–376.
- [44] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation*, Dec. 2001, pp. 722–727.
- [45] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.
- [46] Z. Koldovský and F. Nesta, "Performance analysis of source image estimators in blind source separation," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4166–4176, Aug. 2017.
- [47] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.
- [48] Z. Koldovský, P. Tichavský, and N. Ono, "Orthogonally-constrained extraction of independent non-Gaussian component from non-Gaussian background without ICA," in *Latent Variable Anal. Signal Separation*. Cham, Switzerland: Springer, 2018, pp. 161–170.
- [49] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. Signal Process.*, vol. 45, no. 7, pp. 1712–1725, Jul. 1997.
- [50] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc. F—Commun., Radar Signal Process.*, vol. 130, no. 1, pp. 11–16, Feb. 1983.
- [51] H. Li and T. Adali, "Complex-valued adaptive signal processing using nonlinear functions," *EURASIP J. Adv. Signal Process.*, vol. 2008, Feb. 2008, Article no. 765615.
- [52] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.
- [53] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Proc. Neural Inf. Process. Syst.*, 1996, pp. 757–763.
- [54] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.
- [55] A. Hyvärinen, "The fixed-point algorithm and maximum likelihood estimation for independent component analysis," *Neural Process. Lett.*, vol. 10, no. 1, pp. 1–5, Aug. 1999.
- [56] D. Lahat and C. Jutten, "Joint independent subspace analysis using second-order statistics," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4891–4904, Sep. 2016.
- [57] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adali, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4399–4410, Sep. 2014.
- [58] Z. Koldovský, *Methods for Independent Component/Vector Extraction*, 2018, Accessed: October 15, 2018. [Online]. Available: <https://asap.ite.tul.cz/downloads/ice/>
- [59] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Syst.*, vol. 10, no. 1, pp. 1–8, Feb. 2000.
- [60] S. C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2007, vol. 2, pp. II–637–II–640.
- [61] F. Nesta and Z. Koldovský, "Supervised independent vector analysis through pilot dependent components," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 536–540.



Zbyněk Koldovský (M'04–SM'15) was born in Jablonec nad Nisou, Czech Republic, in 1979. He received the M.S. and Ph.D. degrees in mathematical modeling from the Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, Czech Republic, in 2002 and 2006, respectively. He is currently an Associate Professor with the Institute of Information Technology and Electronics, Technical University of Liberec, Liberec, Czech Republic, and the Leader of Acoustic Signal Analysis and Processing Group. He is the Vice Dean

for Science, Research and Doctoral Studies at the Faculty of Mechatronics, Informatics and Interdisciplinary Studies. His main research interests include audio signal processing, blind source separation, independent component analysis, and sparse representations.

He has served as a General Co-Chair of the 12th Conference on Latent Variable Analysis and Signal Separation, Liberec, Czech Republic, and as a Technical Co-Chair of the 16th International Workshop on Acoustic Signal Enhancement, Tokyo, Japan.



Petr Tichavský (M'98–SM'04) received the Ph.D. degree in theoretical cybernetics from the Czechoslovak Academy of Sciences, Prague, Czech Republic, in 1992, and the Research Professor degree from the same institution in 2017. He is currently with the Institute of Information Theory and Automation, Czech Academy of Sciences in Prague. He is author and coauthor of research papers in the area of sinusoidal frequency/frequency-rate estimation, adaptive filtering and tracking of time-varying signal parameters, algorithm-independent bounds on achievable performance, sensor array processing, independent component analysis, and blind source separation, and tensor decompositions.

He served as an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS from 2002 to 2004 and as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2005 to 2009 and from 2011 to 2016. From 2008 to 2011 and from 2016 till now he has been a member of the IEEE SPS committee Signal Processing Theory and Methods. He was also the General Co-Chair of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2011 in Prague, Czech Republic.