

LABORATOIRE



INFORMATIQUE, SIGNAUX ET SYSTÈMES
DE SOPHIA ANTIPOLIS
UMR 6070

EXTRACTION DE SOURCES PAR MAXIMISATION DE VARIANCE DANS LES QUEUES D'UNE DISTRIBUTION CONDITIONNELLE

Ronald Phlypo, Vicente Zarzoso, Ignace Lemahieu

Equipe SIGNAL

Rapport de recherche
ISRN I3S/RR-2009-07-FR

Mai 2009

RÉSUMÉ :

Ces travaux présentent une méthode pour l'extraction de signaux basée sur les moments conditionnels d'ordre deux de la sortie d'un filtre. L'estimateur du filtre est dérivé d'un critère de vraisemblance approximatif conditionné sur un indicateur de présence de la source d'intérêt. Le moment conditionnel est une fonction de contraste sous les conditions suivantes: 1) les moments croisés entre le signal source d'intérêt et les autres signaux sources sont nuls 2) le moment conditionnel de la source d'intérêt majore les moments conditionnels des autres combinaisons linéaires. Dans le cas de 2 sources, 2 observations et sous la condition qu'ils n'existent pas de moment d'ordre deux croisés entre notre source d'intérêt et les autres sources, on peut établir les bornes théoriques d'estimation. En outre, ces bornes théoriques et les bornes empiriques coïncident. Les simulations montrent que notre estimateur est robuste par rapport au bruit gaussien additif et que les erreurs induites par des approximations de l'indicateur de présence sont semblables aux erreurs induites par du bruit gaussien. La robustesse par rapport au bruit et des approximations de l'indicateur garantissent une applicabilité étendue.

MOTS CLÉS :

Extraction de sources, estimation, fonctions de contrastes, vraisemblance conditionnelle

ABSTRACT:

This work presents a method for signal extraction based on conditional second-order moments of the output of the extraction filter. The estimator of the filter is derived from an approximate maximum likelihood criterion conditioned on a presence indicator of the source of interest. The conditional moment is shown to be a contrast function under the conditions that (1) all cross-moments of the same order between the source signal of interest and the other source signals are null and (2) that the source of interest has the largest conditional moment among all sources. For the two-source two-observation case, this allows us to derive the theoretical recovery bounds of the contrast when the conditional cross-moment does not vanish. A comparison with empirical results confirms these bounds. Simulations show that the estimator is quite robust to additive Gaussian distributed noise. Also through simulations, we show that the error level induced by a rough approximation of the presence indicator shows a strong similarity with that of additive noise. The robustness, both with respect to noise and to inaccuracies in the prior information about the source presence, guarantees a wide applicability of the proposed method.

KEY WORDS :

source extraction, estimation, contrast function, conditional likelihood

Source Extraction by Maximizing the Variance in the Conditional Distribution Tails

Ronald Phlypo*, *Student Member, IEEE*, Vicente Zarzoso, *Member, IEEE* and Ignace Lemahieu, *Senior Member, IEEE*

Abstract—This work presents a method for signal extraction based on conditional second-order moments of the output of the extraction filter. The estimator of the filter is derived from an approximate maximum likelihood criterion conditioned on a presence indicator of the source of interest. The conditional moment is shown to be a contrast function under the conditions that (1) all cross-moments of the same order between the source signal of interest and the other source signals are null and (2) that the source of interest has the largest conditional moment among all sources. For the two-source two-observation case, this allows us to derive the theoretical recovery bounds of the contrast when the conditional cross-moment does not vanish. A comparison with empirical results confirms these bounds. Simulations show that the estimator is quite robust to additive Gaussian distributed noise. Also through simulations, we show that the error level induced by a rough approximation of the presence indicator shows a strong similarity with that of additive noise. The robustness, both with respect to noise and to inaccuracies in the prior information about the source presence, guarantees a wide applicability of the proposed method.

Index Terms—Source Extraction, Estimation, Contrast Functions, Conditional Likelihood.

I. INTRODUCTION

Blind source separation (BSS) aims to recover source signals when only a mixture of them is observed on a sensor array. Recovering the sources means that we have to inverse an estimate of the mixing system, or, more general, estimate an unmixing system, since the mixing system cannot always be inverted (as is the case in under-determined systems, but these will not be treated in this paper). An application of this unmixing system to the observations then yields an output that is an estimate of the source signal. The term blind points out that neither the sources nor the mixing system are explicitly accessible and as such it is impossible to use a distance measure between the sources and the output for the separation as is the case, e.g., in the Wiener filter [1].

In the past two decades, the topic of blind separation has received growing interest. Specially since the introduction of the quite natural model of independent sources, which seems to be an appropriate model for communications and biomedical signal analysis, to give a few examples. The algorithmic and theoretic development of blind source separation under the aforementioned model is now widely known as Independent

Component Analysis (ICA) [2], [3]. The algorithms for ICA are based on the optimisation of a cost function, or contrast, that imposes a measure for independence on the outputs of the unmixing system. It has been shown that the optimisation of any such measure based on the independence of the outputs is sufficient to solve the separation of the observations into the independent sources, up to the inherent ambiguities of scaling, source permutation (order) and phase [4].

Since these ambiguities are waveform preserving, they are generally of no detrimental effect on the separation and hence admissible. However, one might imagine some practical situation where only a single source or a well-defined subset of the sources is of interest and thus the permutation ambiguity is no longer accepted. A straightforward solution would be based on a divide-and-conquer strategy, where the sources are first separated, followed by a selection procedure. However, the total separation would present a non-negligible computational overload, especially when large datasets are considered.

An alternative to the joint separation and a posteriori selection can be found in the BSS algorithms based on a source-by-source extraction [5], [6]. This iterative approach to source separation estimates one of the sources at each iteration. The estimation is followed by a subtraction of the estimate's space from the initial space, as such reducing the dimension of the observation space by one. This method is also known as deflation [5]. It can be shown that the objective functions that are valuable for single source extraction have to be (implicitly) related to the negentropy of the output [7]. Theoretically, the extraction order of the sources can be fixed, based upon their stochastic properties [8], allowing for the more informative sources (higher negentropy) to be extracted first. Nevertheless, this does not improve the major drawback of deflation based algorithms, namely the propagation of the estimation error [5]. And thus, this might be insufficient in some applications, because the specific source of interest is not necessarily the most informative one, appearing only late in the deflation process and accumulating the errors made in the previous iterations.

It is clear from the above that we can not resolve the permutation ambiguity without adding some discriminating information - other than negentropy - about the source of interest into the source extraction objective function. However, the prior information used to discriminate our source of interest from the other sources, should be kept to a minimum if we want to keep the source extraction maximally blind. This is the aim of the class of constrained ICA (cICA) algorithms [9], [10]. The cICA approach puts a constraint on the solution space of the maximum negentropy objective

R. Phlypo and I. Lemahieu are with MEDISIP/IBBT, UGent - Heymans Institute Block B, 185 De Pintelaan, 9000 Ghent, Belgium. E-mail: {ronald.phlypo,ignace.lemahieu}@ugent.be

V. Zarzoso is with UNSA/CNRS, Les Algorithmes - Euclide B, B.P. 121, 2000 Route des Lucioles, 06903 Sophia Antipolis Cedex, France. E-mail: zarzoso@i3s.unice.fr

function, by introducing a penalising term, generally based on a maximally admissible distance measure between the output and a reference signal. In contrast to the solution obtained by minimising the squared error between the filter output and the reference signal (the basis to the Wiener filter [1]), the solution to cICA is (at least theoretically) not the output that has a minimum distance to the reference, but rather the solution that has maximal negentropy among the admissible solutions (under the constraint). A closely related algorithm is BSS with a reference (BSSR) [11], based on the higher order dependencies between the output signal and a reference. The proposed algorithm differs with the Wiener filter mainly in the distance measure that is used. Because higher order moments are considered, BSSR offers a better performance when the reference signal has relatively few non-zero values [11]. Closely related are the Quadratic Higher Order Criteria (QHOC) [12], [13], although the reference signal is no longer held fixed, but instead iteratively derived from an initial random filtering of the observations. This makes the method less suitable for the purpose of single source extraction, where only a single source of interest is envisaged, although we will find that it is closely related to the BSSR method, once the reference is held fixed.

In this paper we make use of the conditional moments of the filter output, rather than using an explicit reference signal. Although this approach to the problem differs from that of reference based filtering, we are able to show that for certain well defined cases, the above algorithms of cICA, BSSR, QHOC and Wiener filtering can be related to the theory presented herein.

The paper begins with an introduction on the signal model and the notational conventions in Section II. We then provide the theoretical aspects of the framework and present the MaxViT method in Section III. Section IV places the presented method in perspective with respect to some competing models and algorithms found in literature and we show that under certain conditions, some explicit or implicit relations can be shown between these alternative methods and the presented method. Because of their similarities, these algorithms and models will be used in comparison studies in Section V after the performance bounds and some properties/characteristics of our method are examined. This will be followed by a discussion in Section VI to conclude with a summary in Section VII.

II. SIGNAL MODEL AND NOTATION

A. Notational Conventions

Scalar variables, column vectors and matrices are respectively given by lowercase lightface (u), lowercase boldface (\mathbf{u}) and uppercase boldface (\mathbf{U}) characters. Consistency of the notations then requires the j -th entry of \mathbf{u} to be denoted by u_j . The probability density function (p.d.f.) associated to the random variable u will be denoted by p_u and the association is denoted as $u \sim p_u$. Realizations of random variables or vectors are respectively given as scalars or vectors with an (arbitrary) indexing to identify the samples, e.g., $\mathbf{u}[k]$ stands for a sample of \mathbf{u} , referenced to by the index k . Also, let

constants be given as uppercase lightface characters (U), the set of real numbers as \mathbb{R} and sets by calligraphic uppercase characters, whose cardinal number is $\#(\mathcal{U})$. A set of K realisations from the random vector \mathbf{u} (a population) is then defined as $\mathcal{U} = \{\mathbf{u}[k] \mid \mathbf{u} \sim p_{\mathbf{u}}, k = 1, 2, \dots, K\}$ and will be referred to by the short-hand notation $\{\mathbf{u}\}_K$, although we will commonly make misuse of the notation and drop K as well as the accolades.

Furthermore, the mathematical expectation of a function f with respect to \mathbf{u} defined as $\int p_{\mathbf{u}}(\mathbf{x})f(\mathbf{x})d\mathbf{x}$ will be denoted by $\mathbb{E}\{f(\mathbf{u})\}$. When conditional on a function g of v it is defined as $\int p_{\mathbf{u}}(\mathbf{x}|g(v))f(\mathbf{x})d\mathbf{x}$ and will be denoted by $\mathbb{E}_{g(v)}\{f(\mathbf{u})\}$. Only if the function g would not be clear from the context, we will denote the conditional expectation by $\mathbb{E}_{g(v)}\{f(\mathbf{u})\}$. Finally, the transpose of a vector \mathbf{u} is written as \mathbf{u}^T .

B. Signal Model

In this work, we use the generative linear model, where an M -dimensional random vector \mathbf{y} can be linked to the underlying N -dimensional random source vector \mathbf{s} (generally $M \geq N$), through the instantaneous linear relation

$$\mathbf{y} = \mathbf{A}\mathbf{s} , \quad (1)$$

which for a limited population can be expressed as $\mathbf{y}[k] = \mathbf{A}\mathbf{s}[k], \forall k = 1 \dots K$. We assume that all random variables are zero-mean, without loss of generality.

III. METHODS

A. Contrast Functions

The goal of blind source separation is to estimate the set $\{\mathbf{s}\}_K$ in the model of Eq. (1) when only having access to a limited set of observations $\{\mathbf{y}\}_K$. *Blind* means that one has neither access to the source vector \mathbf{s} , nor to the mixing matrix \mathbf{A} . Since only a population sample $\{\mathbf{y}\}_K$ of \mathbf{y} is accessible, we need a measure to evaluate the accuracy of the possible estimations of \mathbf{s} , without referencing to $\{\mathbf{s}\}_K$. If we suppose that estimates of \mathbf{s} are restricted to a linear transformation \mathbf{H}^T applied on \mathbf{y} then we have the following relations: $\mathbf{x} = \mathbf{H}^T\mathbf{y} = \mathbf{H}^T\mathbf{A}\mathbf{s} = \mathbf{G}^T\mathbf{s}$. Blind separation of \mathbf{y} into the sources \mathbf{s} can now be done by calculating an appropriate measure of \mathbf{x} . If we choose this measure to be independence (or any approximation thereof), this leads to the contrast functions for ICA [3]. These contrasts can be seen as an extension to the instantaneous linear (generative) model of the objective functions for minimum entropy deconvolution introduced in [14]. It states that for the linear model of Eq. (1), any function $\Psi_{ICA}(\mathbf{x})$ fulfilling the properties

(P1) Invariance

$$\mathbf{B} = \Lambda\mathbf{P}^T \Rightarrow \Psi_{ICA}(\mathbf{x}) = \Psi_{ICA}(\mathbf{B}^T\mathbf{x}) ,$$

(P2) Domination

$$\Psi_{ICA}(\mathbf{s}) \geq \Psi_{ICA}(\mathbf{B}^T\mathbf{s}), \forall \mathbf{B} \text{ and}$$

(P3) Discrimination

$$\Psi_{ICA}(\mathbf{s}) = \Psi_{ICA}(\mathbf{B}^T\mathbf{s}) \Leftrightarrow \mathbf{G} = \Lambda\mathbf{\Pi}^T ,$$

(Λ and $\mathbf{\Pi}$ respectively a diagonal scaling matrix and a permutation matrix) is a contrast function for ICA and can be used to solve the separation problem through its maximisation. In

other words, $\mathbf{x} = [\arg \max_{\mathbf{H}} \Psi_{ICA}(\mathbf{H}^T \mathbf{y})]^T \mathbf{y}$ is an estimation of the source vector. Since the model is based on independent sources, it is easy to verify that permutation and scaling are indeed the inherent indeterminacies of the model.

The same strategy also applies for source extraction, where only a single source, say s_j , is of interest. The above properties then translate into:

- (P1') **Invariance**
 $\lambda \in \mathbb{R} \setminus \{0\} \Rightarrow \Psi(x_j) = \Psi(\lambda x_j)$,
- (P2') **Domination**
 $\Psi(s_j) \geq \Psi(\mathbf{b}^T \mathbf{s}), \forall \mathbf{b}$ and
- (P3') **Discrimination**
 $\Psi(s_j) = \Psi(\mathbf{b}^T \mathbf{s}) \Leftrightarrow \mathbf{b} = \Lambda_j$,

where Λ_j is the j -th column of a diagonal scaling matrix Λ . It should be clear that the permutation ambiguity, although inherent to the separation problem, does no longer exists with the above definition of contrast functions for source extraction once j has been fixed. This is in contrast with previous definitions of contrast functions for source extraction (see e.g. [7], [15], [13]), where the contrast functions have been defined keeping in mind the consecutive extraction of sources for solving the complete separation problem. It is clear that in the latter case, the order in which the sources are extracted is of no importance (cf. the permutation ambiguity in the above definition of contrast functions for ICA).

Fixing j also means that some marginal information is required to distinguish s_j from $s_i, \forall i \neq j$. In this work we propose to use a presence indicator, showing whether or not the sample $s_j[k]$ has been drawn from the tails of its distribution p_{s_j} .

B. From Maximum Likelihood to a Presence Indicator

Consider the model of Eq. (1). Having perfect knowledge of the observations (i.e. the distribution function $p_{\mathbf{y}}$ is known, or can be estimated from the samples \mathbf{y}), the expected observation log-likelihood of \mathbf{A} (and \mathbf{s}) is defined as

$$\begin{aligned} \mathcal{L}(\mathbf{A} | \mathbf{y}) &= \int_{\mathbb{R}^N} p_{\mathbf{y}}(\mathbf{u}) \log p_{\mathbf{y}}(\mathbf{u} | \mathbf{A}, \mathbf{s}) d\mathbf{u} \\ &= \int_{\mathbb{R}^N} |\Delta_{\mathbf{H}}| p_{\mathbf{x}}(\mathbf{u}) \log |\Delta_{\mathbf{H}}| p_{\mathbf{x}}(\mathbf{u} | \mathbf{A}, \mathbf{s}) d\mathbf{u} \\ &= |\Delta_{\mathbf{H}}| \left[\log |\Delta_{\mathbf{H}}| + \int_{\mathbb{R}^N} p_{\mathbf{x}}(\mathbf{u}) \log p_{\mathbf{s}}(\mathbf{u}) \right] \\ &= \int_{\mathbb{R}^N} p_{\mathbf{x}}(\mathbf{u}) \log p_{\mathbf{s}}(\mathbf{u}) , \end{aligned}$$

where $\mathbf{x} = \mathbf{H}^T \mathbf{y} = \mathbf{H}^T \mathbf{A} \mathbf{s}$ is the filter output, $\Delta_{\mathbf{H}} = \det \mathbf{H}$ and we assumed the identities $|\det \mathbf{A}| = |\det \mathbf{H}| = 1$. Remark that these identities can always be fulfilled, e.g. by putting $\mathbf{H} \leftarrow |\Delta_{\mathbf{H}}|^{-1/N} \mathbf{H}$ and $\mathbf{x} \leftarrow |\Delta_{\mathbf{H}}|^{1/N} \mathbf{x}$. Let us furthermore define $\mathbf{H}^T = \mathbf{A}^{-1}$, the maximum likelihood solution is then given as $\mathbf{H} = \arg \max_{\mathbf{H}} \mathcal{L}(\mathbf{H} | \mathbf{y})$.

If we furthermore assume that s_j is independently distributed with respect to $\tilde{\mathbf{s}} = [s_1 \dots s_{j-1}, s_{j+1} \dots s_N]$, the

likelihood becomes

$$\mathcal{L}(\mathbf{H} | \mathbf{y}) = \int_{\mathbb{R}} p_{x_j}(u) \log p_{s_j}(u) du + \int_{\mathbb{R}^{N-1}} p_{\tilde{\mathbf{x}}}(\mathbf{u}) \log p_{\tilde{\mathbf{s}}}(\mathbf{u}) d\mathbf{u} , \quad (2)$$

where $\tilde{\mathbf{x}}$ is defined analogously as $\tilde{\mathbf{s}}$.

With the (log-)likelihood function defined as in Eq. (2), p_{s_j} and $p_{\tilde{\mathbf{s}}}$ have to be available (e.g., through a parametrisation). However, since we are only interested in s_j , we can consider the latter as a nuisance parameter and marginalise our log-likelihood over $p_{\tilde{\mathbf{s}}}$, yielding

$$\mathcal{L}(\mathbf{h}_j | \mathbf{y}) = \int_{\mathbb{R}} p_{x_j}(u) \log p_{s_j}(u) du , \quad (3)$$

however, we still need p_{s_j} , which is rarely available. To circumvent this drawback, we opt for a minimal parametrisation of a distribution p_{s_j} by dividing its support set \mathcal{S} (generally \mathbb{R}) into two half spaces $\mathcal{B} = \{u | |u| > C\}$ and $\bar{\mathcal{B}} = \mathcal{S} \setminus \mathcal{B}$ and define the probabilities

$$\begin{aligned} P_{s_j}(\mathcal{B}) &= \int_{\mathcal{B}} p_{s_j}(u) du , \\ P_{s_j}(\bar{\mathcal{B}}) &= \int_{\bar{\mathcal{B}}} p_{s_j}(u) du = 1 - P_{s_j}(\mathcal{B}) . \end{aligned}$$

This is also defined as a Bernoulli distribution associated to our variable s_j for the events $s_j \in \mathcal{B}$ and $s_j \notin \mathcal{B}$ and brings us to

$$\mathcal{L}(\mathbf{h}_j | \mathbf{y}) = P_{x_j}(\mathcal{B}) \log P_{s_j}(\mathcal{B}) + P_{x_j}(\bar{\mathcal{B}}) \log P_{s_j}(\bar{\mathcal{B}}) .$$

We are still left with the parameter $P_{s_j}(\mathcal{B})$. To overcome this, we can define the conditional marginal likelihood function

$$\begin{aligned} \mathcal{L}(\mathbf{h}_j | \mathbf{y}, \mathcal{I}_{s_j}) &= P_{x_j}(\mathcal{B} | \mathcal{I}_{s_j}) \log P_{s_j}(\mathcal{B} | \mathcal{I}_{s_j}) \dots \\ &\quad + P_{x_j}(\bar{\mathcal{B}} | \mathcal{I}_{s_j}) \log P_{s_j}(\bar{\mathcal{B}} | \mathcal{I}_{s_j}) \end{aligned} \quad (4)$$

where we denoted by \mathcal{I}_{s_j} the event $|s_j| > C$ (i.e. $s_j \in \mathcal{B}$), which is a presence indicator for s_j with respect to a threshold C and where $P_{s_j}(\mathcal{B} | \mathcal{I}_{s_j}) = 1$ and $P_{s_j}(\bar{\mathcal{B}} | \mathcal{I}_{s_j}) = 0$. It is clear that this conditional likelihood heavily penalizes estimates $x = \mathbf{h}^T \mathbf{y}$, which are not conditionally distributed as the sources, or for which $P_{x_j}(\bar{\mathcal{B}} | \mathcal{I}_{s_j}) > 0$. If we denote by $\mathcal{C}_u = \{k | |u[k]| > C\}$ a set of conditional sample indexes, we have that the likelihood for a population sample $\{\mathbf{x}\}_K$ of x strongly penalises the condition $\mathcal{C}_{s_j} \setminus \mathcal{C}_x \neq \emptyset$

In the sequel, let us define $x = x_j$ or $\{x_j\}_K$ and $\mathbf{h} = \mathbf{h}_j$ as well as $\mathcal{L}(x) = \mathcal{L}(\mathbf{h}_j | \mathbf{y}, \mathcal{I}_{s_j})$ for notational convenience.

Let us now verify whether $\mathcal{L}(x)$ can be used as a contrast function for the extraction of s_j from a mixture \mathbf{y} by examining if it fulfils the properties (P1')-(P3').

- (P1') By replacing the condition $|u| > C$ by $|u| > C_u$, with $C_u = C \sigma_u$ in the above definitions (σ_u^2 being the second order moment of u), $\mathcal{L}(x)$ meets property (P1').
- (P2') It is easy to see that $\mathcal{L}(s_j) = 0$, and since we have $\mathcal{L}(x) \in \{0, -\infty\}$, property (P2') is fulfilled for $\mathcal{L}(x)$.
- (P3') The discrimination property requires that $\mathcal{L}(x) = 0$ if and **only** if $x = \lambda s_j$, with $\lambda \in \mathbb{R} \setminus \{0\}$ a scaling factor. This condition should be investigated in more detail, and

we content ourselves by proving that $\mathcal{L}(x)$ fulfils this property for a specific source model.

Proposition 1: $\mathcal{L}(x)$ is a contrast for the source model \mathbf{s} where s_j is independently distributed with respect to $\tilde{\mathbf{s}} = [s_1 \dots s_{j-1}, s_{j+1} \dots s_N]^T$ and this for any $C > 0$.

For a proof, we refer the reader to the appendix.

Now we know that $\mathcal{L}(x)$ is a contrast under the above condition, \hat{s}_j can be found by maximising the conditional marginal likelihood from Eq. (4). Because \hat{s}_j is a singular point and $\mathcal{L}(x) \in \{0, -\infty\}$, the maximisation of $\mathcal{L}(x)$ is equivalent to an exhaustive search over $\mathcal{H} = \{\mathbf{h}\}$ and thus NP hard.

C. Relaxation of the Binary Likelihood Function

To assure the optimisation of the above likelihood function is no longer NP hard, we can replace the conditional probability density function in Eq. (4) by a smooth function. Since in the above conditional log-likelihood, the probability $P_x(\bar{\mathcal{B}} | \mathcal{I}_{s_j})$ was heavily penalised by the term $\log P_{s_j}(\bar{\mathcal{B}} | \mathcal{I}_{s_j})$ in comparison with $P_x(\mathcal{B} | \mathcal{I}_{s_j})$ whose penalising term is $\log P_{s_j}(\mathcal{B} | \mathcal{I}_{s_j})$. Any approximation of the conditional probability should thus share this penalisation, which is equivalent to

$$\frac{P_{s_j}(\bar{\mathcal{B}} | \mathcal{I}_{s_j})}{P_{s_j}(\mathcal{B} | \mathcal{I}_{s_j})} \rightarrow 0 \quad (5)$$

Let us consider a candidate conditional probability

$$p_{s_j}(u | \mathcal{I}_{s_j}, \beta) = \begin{cases} C_1 e^{\beta u^2 / \sigma_{s_j}^2} & \text{if } |x| \leq D \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where the proportionality constant C_1 is the appropriate integration constant and $\sigma_{s_j}^2$ guarantees the invariance with respect to scaling (**P1**).

The above conditional p.d.f. is not necessarily restrictive, because in a practical situation we can always find a D' sufficiently large such that $p_{s_i}(|s_i| > D') \approx 0, \forall i$. Moreover, the definition of Eq. (6) always allows us to choose such a D' together with a value for C such that the condition in Eq. (5) is met. For the generative model, we have $x = \mathbf{g}^T \mathbf{s}$ and thus by taking $\max_i |g_i| < \infty$ it follows that $p_x(|x| > ND' \max_i |g_i|) \approx 0$, thus allowing for any value $D \geq ND' \max_i |g_i|$.

With the above conditional p.d.f. from Eq. (6), the conditional marginalised log-likelihood of the estimate x can be written as

$$\begin{aligned} \mathcal{L}(x) &= \int p_x(u | \mathcal{I}_{s_j}) (\beta u^2 / \sigma_{s_j}^2 + \log C_1) du \\ &= \log C_1 + \beta \int p_x(u | \mathcal{I}_{s_j}) u^2 / \sigma_{s_j}^2 du. \end{aligned} \quad (7)$$

This functional has some attractive algebraic properties, as we will see next.

D. A Contrast with an Algebraic Optimum

The maximisation of $\mathcal{L}(x)$ in Eq. (7) is equivalent to the maximisation of $\Psi(x) = \mathbb{E}_{s_j}\{x^2 / \mathbb{E}\{x^2\}\} = \mathbb{E}_{s_j}\{x^2\} / \mathbb{E}\{x^2\}$, where we use $\mathbb{E}\{x^2\}$ as an estimator for $\sigma_{s_j}^2$

in the maximum likelihood sense. Introducing the shorthand notations $\Phi_{\mathbf{u}}^{s_j} = \mathbb{E}_{s_j}\{\mathbf{u}\mathbf{u}^T\}$ and $\Phi_{\mathbf{u}} = \mathbb{E}\{\mathbf{u}\mathbf{u}^T\}$, we can write

$$\Psi(x) = \frac{\Phi_x^{s_j}}{\Phi_x} = \frac{\mathbf{h}^T \Phi_{\mathbf{y}}^{s_j} \mathbf{h}}{\mathbf{h}^T \Phi_{\mathbf{y}} \mathbf{h}}. \quad (8)$$

This is a generalised Rayleigh quotient, and its maximisation has as an algebraic solution; see e.g., [16, Sec. 8.7.1].

The maximisation of Eq. (8) can be done through the eigenvalue decomposition of $\Phi_{\mathbf{y}}^{-1} \Phi_{\mathbf{y}}^{s_j}$ (whenever $\Phi_{\mathbf{y}}$ is invertible) and choosing the major eigenvector/eigenvalue pair \mathbf{q}, λ for which

$$\Phi_{\mathbf{y}}^{-1} \Phi_{\mathbf{y}}^{s_j} \mathbf{q} = \lambda \mathbf{q}. \quad (9)$$

λ is the maximum function value and \mathbf{q} is the estimate of \mathbf{h} from which we obtain $x = \hat{s}_j = \mathbf{h}^T \mathbf{y}$.

We have already seen that the log-likelihood as defined in Eq. (4) is a contrast function for a source s_j independently distributed with respect to $s_i, i \neq j$. However, since $\Psi(x)$ in Eq. (8) is an approximation thereof, we need to investigate under what conditions the above approximate likelihood may be considered as a contrast function.

Proposition 2:

$$\Psi(x) = \frac{\Phi_x^{s_j}}{\Phi_x} \text{ subject to } x = \mathbf{h}^T \mathbf{y} \quad (10)$$

is a contrast for the extraction of s_j under the following sufficient conditions $\forall i \neq j$:

- (C1) $\mathbb{E}\{s_j s_i\} = 0$;
- (C2) $\mathbb{E}_{s_j}\{s_j s_i\} = 0$;
- (C3) $\mathbb{E}_{s_j}\{s_j^2\} / \mathbb{E}\{s_j^2\} > \mathbb{E}_{s_j}\{s_i^2\} / \mathbb{E}\{s_i^2\}$.

For the proof, we refer the reader to the appendix.

Remark that the statistical independence of s_j with respect to $\tilde{\mathbf{s}}$ is no longer a necessary condition, and that this condition has been relaxed to second-order independence (decorrelation) only.

IV. CONNECTION TO OTHER METHODS

While the starting point of our method is quite different than those of the methods that will be discussed below, certain connections with some methods in literature can be made. We insist on clarifying the coherences between those methods and the proposed MaxViT method before the presentation of the results as to motivate our choice of algorithms used in later comparisons.

A. ICA

In most practical cases, the mutual independence of the sources is an acceptable prior, which makes ICA one of the most popular source separation algorithms nowadays [17], [18], [19], [3]. We prove next that the independence of the sources is a sufficient condition to be recovered by the approximate maximum likelihood estimator of MaxViT, under the assumption that the conditional covariance can be calculated, i.e., the set \mathcal{C}_{s_j} should be available.

Since the independence of the entries already assures that (C1) and (C2) are met, we are left to show the plausibility of (C3) under the independence assumption. Independence

means that $p(u|s_j) = p(u), \forall u \neq f(s_j)$, where $f(\cdot)$ can be any function. We thus have $\mathbb{E}_{s_j}\{s_i^2\} = \mathbb{E}\{s_i^2\}, \forall i \neq j$ and $\mathbb{E}_{s_j}\{s_j^2\} > \mathbb{E}\{s_j^2\}$, where the last inequality is proven in the appendix. In addition, the results obtained in the appendix, allow us to alter the condition in Eq. (5) to $\int_{\bar{\mathcal{B}}} p_{s_j}(u)g(u)du \rightarrow 0$, which is a condition on the function $g(u) = \log \hat{p}_{s_j}(u|\mathcal{I}_{s_j})$, but now in relation to $p_{s_j}(u)$.

B. Reference-based Filtering

When a reference signal is available for the extraction of a source, one can use extraction filters such as obtained, amongst others, via the optimal Wiener filter estimate [1] or via Blind Source Separation with a Reference (BSSR) [11]. In this section we show that by choosing the right reference for the Wiener filter or the BSSR method, we obtain the same result as with the approximate maximum likelihood estimator of MaxViT under certain conditions.

Consider first the Wiener filter $\mathbf{h}_W = \mathbb{E}\{\mathbf{y}\mathbf{y}^T\}^{-1}\mathbb{E}\{\mathbf{y}r\}$, where r is the reference signal. Taking as a reference $r = s_j$, we have $\mathbf{h}_W = \Phi_{\mathbf{y}}^{-1}\mathbb{E}\{\mathbf{y}s_j\}$ and the variance of the output $x_W = \mathbf{h}_W^T \mathbf{y}$ is

$$\begin{aligned} \Phi_{x_W} &= \mathbb{E}\{s_j \mathbf{y}^T\} \Phi_{\mathbf{y}}^{-1} \mathbb{E}\{\mathbf{y} s_j\} \\ &= \mathbb{E}\{s_j \mathbf{s}^T\} \mathbf{A}^T (\mathbf{A} \Phi_{\mathbf{s}} \mathbf{A}^T)^{-1} \mathbf{A} \mathbb{E}\{s_j \mathbf{s}\} \\ &= \mathbb{E}\{s_j^2\}^2 [\Phi_{\mathbf{s}}^{-1}]_{jj} , \end{aligned} \quad (11)$$

where the last equality follows from **(C1)**, from which follows that $[\Phi_{\mathbf{s}}^{-1}]_{jj} = (\det \Phi_{\mathbf{s}})^{-1} (\det \Phi_{\bar{\mathbf{s}}}) = (\det \Phi_{\bar{\mathbf{s}}})^{-1} [\Phi_{\mathbf{s}}]_{jj}^{-1} (\det \Phi_{\bar{\mathbf{s}}}) = [\Phi_{\mathbf{s}}]_{jj}^{-1} = \mathbb{E}\{s_j^2\}^{-1}$. The conditional variance is analogously given by

$$\Phi_{x_W}^{s_j} = \mathbb{E}\{s_j^2\}^2 [(\Phi_{\mathbf{s}^{s_j}})^{-1}]_{jj} , \quad (12)$$

where $[(\Phi_{\mathbf{s}^{s_j}})^{-1}]_{jj} = [(\Phi_{\mathbf{s}^{s_j}})]_{jj}^{-1} = \mathbb{E}_{s_j}\{s_j^2\}^{-1}$ if **(C2)** is fulfilled.

The value of the solution to the Wiener filter in the contrast function can be given by combining Eqs (11) and (12) and putting them into Eq. (8), yielding

$$\Psi(x_W) = \frac{\Phi_{x_W}^{s_j}}{\Phi_{x_W}} = \frac{\Phi_{s_j}}{\Phi_{s_j}^{s_j}} \leq 1 \leq \frac{\Phi_{s_j}^{s_j}}{\Phi_{s_j}} , \quad (13)$$

with equalities if and only if $\Phi_{s_j} = \Phi_{s_j}^{s_j}$, or $P_{s_j}(\bar{\mathcal{B}}) = 0$. Unfortunately, we then no longer have the dominance of the source s_j in the contrast function since all sources now satisfy $\Phi_{s_i}^{s_j} / \Phi_{s_i} = 1$.

For the BSSR method we have the more general objective function (defined for real variables):

$$\phi_{BSSR}^{(p)}(\mathbf{h}) = \frac{1}{2n} \mathbb{E}\{(\mathbf{h}^T \mathbf{y})^{2p} r^{2p}\} - \frac{\lambda}{2} (\mathbf{h}^T \mathbf{h} - 1) ,$$

where r is the reference signal and $2p$ the order. This may be seen as a special case of the QHOC [12], [13], where we have (in the real case)

$$\phi_{QHOC}^{(R)}(\mathbf{h}) = \mathbf{h}^T \text{Cum}\{y^2 r^{R-2}\} \mathbf{h}, \text{ subject to } \mathbf{h}^T \Phi_{\mathbf{y}} \mathbf{h} = 1 ,$$

although the reference is held fixed and is no longer iteratively updated. An iterative fixed-point algorithm was proposed in [11] to optimize this function. Nevertheless, algebraic solutions exist at orders $p = 1/2$ and $p = 2$. At $p = 1/2$

the BSSR cost function accepts the closed-form solution $\mathbf{q} = \mathbb{E}\{\mathbf{y}r\}$; this is the optimal Wiener filter associated with reference signal r if the observations \mathbf{y} are spatially white ($\mathbb{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{I}_m$). At order $p = 1$, the cost function can also be solved algebraically [13]; indeed, \mathbf{h} is then given by the dominant eigenvector of the reference-weighted covariance matrix $\mathbb{E}\{\mathbf{y}\mathbf{y}^T r^2\}$. The BSSR method (and thus also QHOC) at order $2p = 2$ (respectively any pair order R) seems equivalent to the MaxViT method when using a binary reference signal

$$r[k] = \begin{cases} \text{sign}(s_j[k]) & \text{if } |s_j[k]| > C \\ 0 & \text{otherwise} \end{cases} , \quad (14)$$

and under the condition of a spatially white observation vector \mathbf{y} for BSSR. The latter is not explicitly required by MaxViT which renders MaxViT less susceptible to the performance bounds imposed by a prewhitening stage [20]. It should also be noted that the BSSR method has been proposed with a specific application in mind, and little research has been done on its convergence and robustness, whereas the QHOC restricted the references to be a linear combination of the observations. In this paper, we make grateful use of the coherence between these methods to show that they are near optimal in the sense that they are derived from a quasi-likelihood. Additionally, we will show the robustness with respect to arbitrary binary references, which we prefer to address as a conditional indicator function, as such providing as a byproduct a study of the robustness of the BSSR method for the aforementioned specific case. Remark that, whenever we will refer to BSSR in the sequel, we refer to the iterative version of [11].

V. RESULTS

A. Estimation Bounds of MaxViT

In this section we establish the error bounds on the estimation of s_j in the model $x = \mathbf{h}^T \mathbf{A} \mathbf{s} = \mathbf{g}^T \mathbf{s}$. This error can be measured through the interference to signal ratio (ISR) defined as

$$\text{ISR} = \frac{\sum_{i \neq j} |g_i|^2}{(n-1)|g_j|^2} , \quad (15)$$

which is a measure for the average interference, and takes the value zero if and only if the extraction filter is the j -th canonical vector.

The filter \mathbf{g} is the product of the dominant generalised eigenvector \mathbf{h} associated to $\mathbf{A} \Phi_{\mathbf{s}^{s_j}} \mathbf{A}^T / \mathbf{A} \Phi_{\mathbf{s}} \mathbf{A}^T$ and \mathbf{A} . Here, we only consider the bias in the estimate of \mathbf{h} as a consequence of the non-vanishing conditional covariance between s_j and $s_i, i \neq j$. The ISR as a function of this covariance can be calculated for $\mathbf{s} \in \mathbb{R}^2$ as:

$$\text{ISR}(\rho, \delta) = \frac{\sqrt{\delta^2 + |\rho|^2} - \delta}{\sqrt{\delta^2 + |\rho|^2} + \delta} \stackrel{\delta \neq 0}{=} \frac{\sqrt{1 + (|\rho|/\delta)^2} - 1}{\sqrt{1 + (|\rho|/\delta)^2} + 1} , \quad (16)$$

where

$$\delta = (\Phi_{s_j}^{s_j} - \Phi_{s_i}^{s_j}) / 2 \quad (17)$$

$$\rho = \mathbb{E}_{s_j}\{s_j s_i\} . \quad (18)$$

The calculations for the value of ISR are given in appendices C and D and the relation between $|\rho|/\delta$ and the theoretic

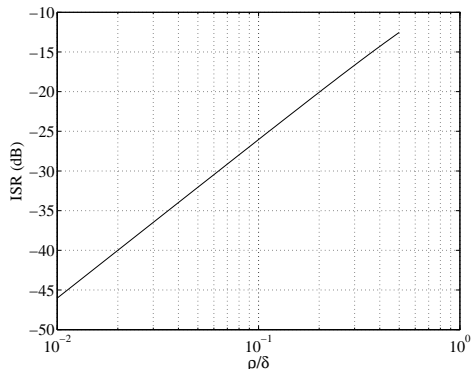


Fig. 1. The theoretical bounds for the value of ISR as a function of the conditional covariance ρ and the conditional variance domination δ .

cal ISR value is given in Fig. 1. We can give an impression of the accuracy of this theoretical measure by comparing it with the obtained ISR as obtained through the relation of Eq. (15). We did this for 1000 Monte Carlo realisations of 2 i.i.d. (respectively Laplacian, Gaussian and uniformly distributed) unit variance, zero-mean source signals of $K = 1000$ samples each observed through an orthonormal mixing matrix \mathbf{A} . With this simulations, we obtained a maximum absolute error of 9.4410^{-16} confirming the accuracy of Eq. (16).

From both Eq. (16) and Fig. 1, we see that $x = \hat{s}_j \approx s_j$ if $|\rho| \ll \delta$ and a good estimation of the source s_j is guaranteed even if $|\rho|$ tends to δ (we still have a theoretical -10dB if $|\rho| = \delta$), which is a reasonable assumption in many practical situations. It can be seen that the smaller the discrimination in the conditional variance becomes in (C3), the more stringent the condition (C2) on $|\rho|$ becomes (and thus automatically also (C1)). Under the condition $|\rho| = 0$, we have already shown that Eq. (10) is a contrast for the separation of s_j from a mixture in Section III-D and, indeed, we obtain $\text{ISR} = 0$ from Eq. (16), no matter the value of δ .

Table I gives the mean fraction of $|\rho|/\delta$ for three different distributions (Uniform, Laplacian and Normal) and for different values of C . Note that the number of sample indexes in the set \mathcal{C}_{s_j} differs according to the chosen distribution, and consequently has a considerable influence on the variance of the statistics $\hat{\mathbb{E}}_{s_j}\{f(x)\} = \sum_{k \in \mathcal{C}_{s_j}} f(x[k]) / \#\mathcal{C}_{s_j}$. Therefore, we decided to use K samples on a basis of K_b , where $K = K_b/\xi$ and $\xi = P_{s_j}(\mathcal{B})$. This brings the number of sample indexes in \mathcal{C}_{s_j} from which $\mathbb{E}_{s_j}\{f(u)\}$ is estimated to an almost equal number, independent of the distribution used. The ISR or the fraction $|\rho|/\delta$ can now directly be compared for a given K_b .

TABLE I

THE FRACTION $|\hat{\rho}|/\hat{\delta}$ FOR DIFFERENT DISTRIBUTIONS AND DIFFERENT VALUES FOR c BASED ON UNIT VARIANCE, ZERO MEAN I.I.D. REALISATIONS AND A BASIS OF $K_b = 10^3$ SAMPLES (SEE TEXT). THE VALUES ARE GIVEN AS MEAN \pm STANDARD DEVIATION.

	Uniform	Normal	Laplace
$c = 1$	0.12 \pm 0.09	0.09 \pm 0.07	0.07 \pm 0.06
$c = \sqrt{2}$	0.12 \pm 0.10	0.09 \pm 0.07	0.07 \pm 0.06
$c = \sqrt{3}$	N/A ¹	0.10 \pm 0.06	0.09 \pm 0.06

The above Eq. (16) is a compact expression for the case $\mathbf{s} \in \mathbb{R}^2$, but for $\mathbf{s} \in \mathbb{R}^n$, $n > 2$ the calculations become more cumbersome. For $n = 3$ we turn to simulations on a synthetic dataset, for which we give the results below.

B. Performance Comparison

To compare the performance of our algorithm with respect to the related algorithms discussed in Section IV, a dataset has been created based on realisations of a source vector $\mathbf{s} \in \mathbb{R}^3$, for which we have $K = 1000$ realizations. The entries in $\{\mathbf{s}\}_{10^3}$ are samples from an i.i.d. unit-variance, zero-mean Laplacian distribution. The so-obtained source signals are then transformed through a unitary matrix \mathbf{A} to the observation space $\mathbf{y} = \mathbf{A}\mathbf{s}$. Without further specifications, we have set $C = 1$ to determine the conditional probabilities.

The algorithms of the Wiener filter and our MaxViT algorithm both have a closed form solution, whilst the ICA algorithm (COM2 [3], without pre-whitening, since we have a unitary mixture) and the BSSR algorithm (taken at higher order $2p = 4$ for the evident reason of avoiding similarity with our MaxViT contrast, see section IV-B) are iterative. The COM2 algorithm has been run over the classical $[1 + \sqrt{n}]$ sweeps over all the signal pairs, which guarantees (although heuristically) its convergence. The BSSR algorithm has either been run until convergence or over 10^3 iterations, whatever has been reached first. Since COM2 provides a separation rather than an extraction, we only retained the output x_i that had the highest correlation with s_1 , the source of interest.

Both BSSR and the Wiener filter can be used with different reference signals. To restrict the wide scope of possibilities, we retain only those references that have a close resemblance with the conditional used for MaxViT, i.e., through the indicator function $\mathcal{I}_{s_1, C}([k]) = 1 \Leftrightarrow k \in \mathcal{C}_{s_j}$. The so obtained reference signal r is then defined as

$$r[k] = \begin{cases} s_j[k] & \text{if } |s_j[k]| > C \\ 0 & \text{otherwise} \end{cases} \quad (19)$$

Derivations of this reference function defined as $b = \text{sign}(r)$ (see also Eq. (14)) or $|b| = |\text{sign}(r)|$ are also used, where we define $\text{sign}(0) = 0$. Similar reference functions have also been proposed, e.g., in [21], [10]. All experiments are evaluated over 1000 Monte Carlo realisations of $\{\mathbf{s}\}_{10^3}$ and \mathbf{A} .

In Table II, we show the mean ISR value as defined in Eq. (15). The ISR is a measure that quantitatively measures the estimation of the filter \mathbf{h} , through an evaluation of $\mathbf{g} = \mathbf{A}^T \mathbf{h}$. Contrary to measures such as Pearson's correlation coefficient, it is an asymptotic evaluation of the interference to signal ratio, and does not make any assumption on the distribution of the error. Table II is organised in such a way, that, reading it from left to right, the information content in the reference signal decreases. The values between brackets are obtained after a rotation of the i.i.d. vector \mathbf{s} by a unitary matrix. This results in decorrelated entries of \mathbf{s} that are no longer guaranteed independent.

¹For $c = \sqrt{3}$, we have $\nu_{s_j}(\mathcal{B}) = 0$ and our method is not applicable (N/A).

TABLE II

ISR AS A MEASURE FOR THE ASYMPTOTIC ACCURACY OF THE SOURCE ESTIMATE FROM A SYNTHETIC DATASET OF 3 I.I.D. LAPLACIAN SOURCES OF $K = 10^3$ SAMPLES FOR DIFFERENT ALGORITHMS AND DIFFERENT INFORMATION FEEDS. VALUES BETWEEN BRACKETS ARE OBTAINED FROM DECORRELATED SOURCES WHICH ARE NOT INDEPENDENT.

	r	b	$ b $	no ref.
MaxViT(1)	.	.	-36.01	.
	.	.	(-28.62)	.
MaxViT($\sqrt{3}$)	.	.	-34.19	.
	.	.	(-28.58)	.
Wiener	-36.61	-35.80	17.31	.
	(-29.37)	(-28.91)	(17.08)	.
BSSR $_{p=2}$	-26.13	-31.91	-31.91	.
	(-25.37)	(-29.30)	(-29.30)	.
COM2	.	.	.	-33.80
	.	.	.	(-24.26)

C. Influence of Additive Noise

We start from the same observations and source signals as defined above. To discard the influence of the parameter quotient $|\rho|/\delta$ on the ISR - see Eq. (16) - we assure that we have $p(|s_i| > c\sigma_{s_i} \mid |s_j| > C\sigma_{s_j}) = p(|s_j| > C\sigma_{s_j} \mid |s_i| > C\sigma_{s_i}) = 0, \forall i \neq j$ by permuting the samples of $\{s_j\}_K$ appropriately. To test the performance of the algorithm under noisy conditions, centred Gaussian noise $\eta \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 \mathbf{I}_3)$ has been added to the observations \mathbf{y} . Since the observations are standardized and the noise is isotropic, the signal to noise ratio (SNR) can be expressed with the simple relation $\text{SNR} = \sigma_\eta^{-2}$. The model reads $\mathbf{y} = \mathbf{A}\mathbf{s} + \eta$ and the estimate of s_j is $x = \mathbf{g}^T \mathbf{s} + \mathbf{g}^T \eta$.

The influence of the SNR on the performance parameter ISR is shown in Fig. 2. Since in the case of additive noise, an accurate estimate of the filter does not guarantee an accurate estimate of the source, we also give the value of $1 - |\hat{\rho}|$, with $\hat{\rho}$ the sample estimate of $\mathbb{E}\{x s_1\} / (\mathbb{E}\{x^2\}^{1/2} \mathbb{E}\{s_1^2\}^{1/2})$. This direct comparison between the source estimate x and the source s_1 can be found in Fig. 3. The comparison of MaxViT has been carried out with respect to the algorithms used in Table II, however, making a selection of reference signals which we judged most useful for comparison. This includes the performance of a Wiener filter and the BSSR method with an unsigned reference $|b|$, adding exactly the same amount of information as is used in MaxViT.

To complete the performance picture, we also add MaxViT with $c = \sqrt{3}$ for comparison. Note that in Fig. 3, the Wiener solution has all of its performance values out of the range used for plotting ($\text{ISR}(\text{Wiener}(|b|)) \in [10, 50] \text{dB}$).

D. Robustness with Respect to the Conditional Set

Assume we no longer have $p(\cdot \mid |s_j| > C\sigma_{s_j})$ but rather $p(\cdot \mid |s_j| + \eta > C\sigma_{s_j})$, where η is a nuisance parameter expressing the uncertainty we have about our initial condition. As before, let us denote by $\mathcal{C}_{s_j} = \{k \mid |s_j[k]| > C\sigma_{s_j}\}$ the conditional set of sample indexes. We can now suppose that the samples are no longer independently drawn but rather

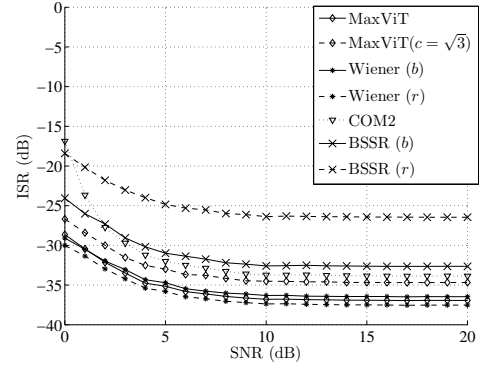


Fig. 2. The source interference ISR (dB) as a function of the signal to noise ratio SNR (dB). The noise is normally distributed additive noise (see text for details). The method is compared with a classical ICA method, the BSSR solution and the solution by a Wiener filter.

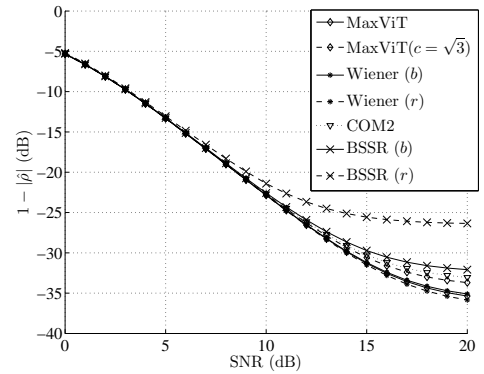


Fig. 3. The 'correlation' $1 - \hat{\rho}(x, s_1)$ (dB) as a function of the signal to noise ratio SNR (dB). The noise is normally distributed zero-mean additive isotropic noise (see text for details). The method is compared with a classical ICA method, the BSSR solution and the solution by a Wiener filter.

conditionally on their index k , which gives us a sample-based conditional $p(u[k] \mid k \in \mathcal{C}_{s_j})$. In what follows we experimentally analyse the robustness of the algorithm with respect to a mismatch of the conditional set \mathcal{C}_{s_j} .

As above, we have $K = 1000$ realisations of three i.i.d. standardised Laplacian sources \mathbf{s} observed through the observations \mathbf{y} through a unitary mixture \mathbf{A} . The samples of s_j have been permuted such that $\forall i \neq j, \mathcal{C}_{s_j} \cap (\bigcup_i \mathcal{C}_{s_i}) = \emptyset$ and thus the source s_j can be estimated since we have $\mathbb{E}\{s_i s_j\} \approx 0$, $\mathbb{E}_{s_j}\{s_i s_j\} \approx 0$ (i.i.d. variables) and $\Phi_{s_j}^{s_j} / \Phi_{s_j} > 1$ (see Section V-A). Remark that we artificially lowered the theoretical ISR estimation bound by permuting the samples and thus augmenting δ . Also, define the following sets of sample indexes:

- $\mathcal{K} = \{k \mid k \in \mathbb{N}, 1 \leq k \leq K\}$
- $\bar{\mathcal{C}}_{s_i} = \mathcal{K} \setminus \mathcal{C}_{s_i}$
- $\mathcal{C}_{ne} = \bigcap_i \bar{\mathcal{C}}_{s_i}$
- $\mathcal{C}_{co,j} = (\bigcup_{i \neq j} \mathcal{C}_{s_i}) \setminus \mathcal{C}_{s_j}$

The latter two sets are respectively the neutral and the conflicting set with respect to s_j .

Consider also the following three set operations:

- Shrinkage(\mathcal{P}_1, n):
 $\mathcal{P}_2 \subseteq \mathcal{P}_1$ with $\#(\mathcal{P}_2) = (1 - n) \times \#(\mathcal{P}_1)$

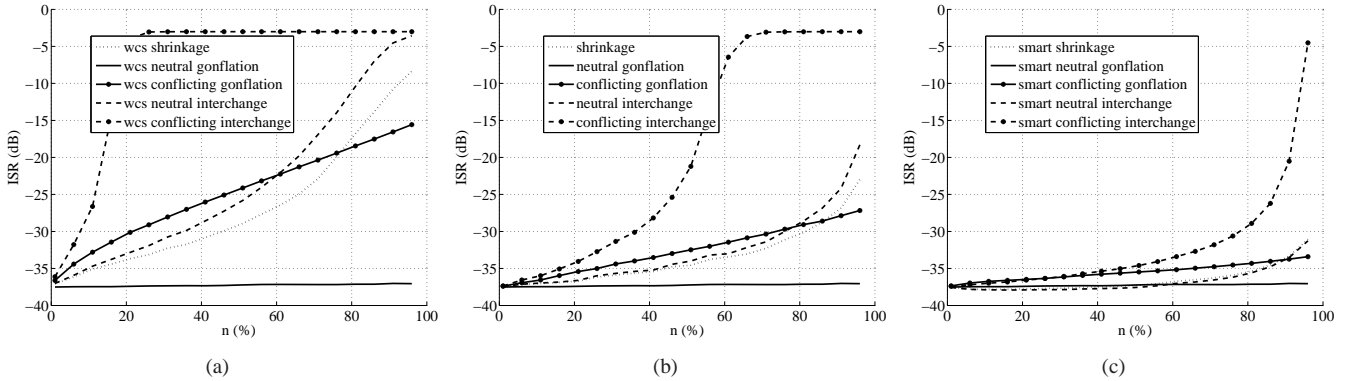


Fig. 4. The effect of a perturbation of the conditional set \mathcal{C}_{s_j} . The effect of the mismatch is measured through the source interference ISR (15) as a function of the relative number of samples n that are affected by the set operations. See text for more details.

- Gonflation($\mathcal{P}_1, \mathcal{Z}, n$):
 $\mathcal{P}_2 = \mathcal{P}_1 \cup \mathcal{Z}$ with $\#(\mathcal{P}_2) = (1 + n) \times \#(\mathcal{P}_1)$
- Interchange(\mathcal{P}_1, n):
 Gonflation($\text{Shrinkage}(\mathcal{P}_1, n), \mathcal{Z}, \frac{n}{1-n}$)

where $\#$ is the cardinal number of the set and n is expressed in percent.

By applying set operations to \mathcal{C}_{s_j} , we obtain an estimate of the perturbation of the conditional probability $p(\cdot \mid |s_j| + \eta > C\sigma_{s_j})$ as has been explained above. The results of this perturbation study can be found in Fig. 4, where we present the results of the above defined set operations on \mathcal{C}_{s_j} . The set \mathcal{Z} is chosen as \mathcal{C}_{ne} or $\mathcal{C}_{co,j}$ for a neutral, respectively a conflicting operation with respect to s_j . The influence of the set perturbation is expressed in terms of the source interference ratio ISR (15).

If the indices of the samples on which the actions are performed form a random subset of \mathcal{C}_{s_j} we obtain the results of Fig. 4(b). Any of the above set operations could be carried out by a more careful selection of the samples of \mathcal{C}_{s_j} and of \mathcal{Z} that determine the set operations. If we shrink the set \mathcal{C}_{s_j} , maintaining the largest entries in \mathcal{C}_{s_j} and/or gonflate \mathcal{C}_{s_j} adding only the smallest entries from \mathcal{C}_{ne} or $\mathcal{C}_{co,j}$, we are reducing the influence to a minimum, since we keep δ in Eq. (16) maximal (while only having relatively small changes of $|\rho|$ for sufficiently low n). Let us call this a smart choice of the subsets, of which the results are given in Fig. 4(c). The inverse is denoted as the worst case scenario and is given in Fig. 4(a).

VI. DISCUSSION

The performance of MaxViT in the noiseless case has shown competitive results with respect to the algorithms used in the comparison Table II. MaxViT even outperforms the reference based algorithms BSSR ($2p = 4$) and the limited support Wiener filter that have access to a larger amount of information (b instead of $|b|$ makes a 1 bit per sample information gain). We also outperform a completely blind algorithm based on higher order statistics (COM2), showing the advantage of using a probability conditioned on the source of interest only. Moreover, the little performance gain that can be obtained by the Wiener filter is at the expense of a highly informative

prior, using the waveform r from Eq. (19), which is generally not available. In an observation environment contaminated by isotropic Gaussian distributed zero-mean i.i.d. noise, the MaxViT estimator shows to be robust, being competitive with the methods used in the comparison, with a slight estimation gain over almost the whole SNR range used in the simulations (Figs 2, 3). The only competitor that outperforms MaxViT is the Wiener estimate, using the source samples on the sample indexes belonging to \mathcal{C}_{s_j} to construct the reference, but as mentioned above, this goes at a cost of a non-negligible information gain in the prior.

We also observe from Table II that the performance of BSSR remains equal, whether a signed or unsigned binary reference is used. This is an immediate consequence of the limitation of the BSSR algorithm to use even powers in the reference ($2p$) [11]. Surprisingly, as can be seen from the same Table II, the BSSR (and we can extrapolate this performance to the QHOC method) algorithm does not yield a significant increase in estimation accuracy with an increase in available information from b to r . This points out that the conditional relative variance may be seen as a sufficient statistic to extract the source s_j from the mixture.

The MaxViT estimator also has been shown to be quite robust to mismatches with respect to the conditional set \mathcal{C}_{s_j} , see Fig. 4(b). This distinguishes our method from other works such as [9], [22], [11], where the estimator is reported to be susceptible to mismatches between the used reference and $s[k]_j$, especially with respect to its phase. Heuristically, BSSR has already been shown to be robust to reference mismatches including time shift and sample omission (i.e., the equivalent of a shrinkage of the set \mathcal{C}_{s_j}) [23]. By placing BSSR in the framework of MaxViT, this can now partly be explained by the robustness of MaxViT to the conditional set \mathcal{C}_{s_j} . This follows from the supposition that BSSR with a binary reference inherits certain properties of MaxViT, whilst being equivalent to MaxViT for $2p = 2$ and reference signal b defined in Section V-B.

The errors induced by the mismatch between the conditional set \mathcal{C}_{s_j} and $\hat{\mathcal{C}}_{s_j}$ added as a prior to the algorithm are comparable to those induced by additional noise as has been suggested in Sec. V-D. However, notice that small errors in the filter estimate induce a smaller error in the filter output than

the additional noise does. This can be deduced from the fact that our filter output can be written as a function of the optimal filter \mathbf{h}^* and a perturbation $\varepsilon_{\mathbf{h}}$ as $x^* + \varepsilon_x = (\mathbf{h}^* + \varepsilon_{\mathbf{h}})^T \mathbf{y}$, whereas in the case of additional noise, the same error in the filter estimate results in $x^* + \varepsilon_{x,2} = (\mathbf{h}^* + \varepsilon_{\mathbf{h}})^T (\mathbf{y} + \eta) = x^* + \varepsilon_x + (\mathbf{h}^* + \varepsilon_{\mathbf{h}})^T \eta$. Thus for the same error in the filter estimate, we would obtain a better estimate of the source if the error is due to the set mismatch only.

Note that despite the use of optimally designed simulations to reduce the fraction $|\rho|/\delta$ and thus to minimise the ISR (by choosing $\mathcal{C}_{s_j} \cap (\bigcup_{i \neq j} \mathcal{C}_{s_i}) = \emptyset$), we may generalise our results to independently distributed sources that have not been corrected, since from Table I we have that the fraction $|\rho|/\delta$ generally remains acceptably small for i.i.d. Laplacian, Gaussian and uniform sources.

A quick overview of the performance of the MaxViT algorithm can be given by evaluating under what conditions we obtain an acceptable ISR of $-30dB$. It follows from Figs 2 and 3 that we accept a signal to noise level no lower than 4dB and (from Figs 4(a)-(c)) a worst case interchange of indexes of \mathcal{C}_{s_j} with \mathcal{C}_{s_i} of up to 7% of $\#\mathcal{C}_{s_j}$. However, in practical situations, an estimation of the set \mathcal{C}_{s_j} is usually done with more care and even when unfortunately chosen, we would randomly interchange samples between sets. This can be done for up to some 30% to 70% of the samples of \mathcal{C}_{s_j} , depending on whether conflicting, respectively neutral sample indexes are added. In practical situations, a set estimate $\hat{\mathcal{C}}_{s_j}$ offering a considerable performance should thus often be available, e.g., by using a threshold on the amplitude of the observations (as in [22]) or based upon prior knowledge of the support in the frequency domain (see e.g., [24]).

The estimation of a source s_j from a set of observations \mathbf{y} can be done for every source in the mixture (approximately) satisfying the sufficient conditions (C1)-(C3) and this whenever an approximation of its conditional set is available. When more than one source is of interest, we propose an iterative estimation without deflation, especially when $\#(\mathcal{C}_{s_i} \cap \mathcal{C}_{s_j})$ is relatively small. Avoiding the subtraction of the projection of \mathbf{y} onto s_j from \mathbf{y} prior to estimating s_i , reduces the possible error propagation from which the deflation approaches suffer.

Also interesting is the similarity between the sparsity pursuit methods (e.g. [25]), where the objective is to have a low approximation error of the observations (with respect to some measure, generally ℓ_2) with only few representative basis functions and our MaxViT model, aiming at minimising the approximation error (through a maximisation of the explained variance of the observations) on a limited amount of samples (the basis functions being Dirac functions). While MaxViT needs a prior knowledge about the presence of s_j (in absolute value with respect to a threshold level \mathcal{C}_{s_j}), sparsity pursuit for multidimensional signals aims at searching a combination of the minimum number of dictionary elements needed to approximate the observations [26], when the mixing matrix \mathbf{A} is known. Combining these two strategies would give a weighted conditional covariance $\mathbb{E}_{s_j}\{\mathbf{y}\mathbf{y}^T\} = \mathbb{E}\{\mathbf{y}\mathbf{D}\mathbf{D}^T\mathbf{y}^T\}$, where \mathbf{D} can be any dictionary (including the ensemble of Dirac functions). Maximising the MaxViT contrast under a maximum sparsity constraint could then be done jointly over

\mathbf{h} and \mathbf{D} , but this should be investigated in more detail.

Within the framework of sparse component analysis, MaxViT - calculating the variance dominance on a subset of the observations - can then also be seen in the category of algorithms based on piecewise linear source separation [26], [27]. The latter has the basic assumption that outside the support of the source of interest, its amplitude is zero or is captured in the background noise with a predefined (low) noise variance. MaxViT has an equivalent assumption on the source presence, namely that it is negligible, except on the *support* \mathcal{C}_{s_j} . This is even more clear when we consider the following slightly adapted MaxViT contrast function,

$$\Psi'(x) = \frac{\mathbb{E}_{s_j}\{x^2\} - \mathbb{E}\{x^2\}}{\mathbb{E}\{x^2\}} = \Psi(x) - 1 .$$

This equation has the same maximisers, but we now have that all other eigenvalues - other than the major eigenvalue - equal 0. The maximisation of $\Psi(x)$ is thus equivalent to a search for a matrix \mathbf{H} that renders the diagonal of $\mathbf{H}^T(\mathbb{E}_{s_j}\{\mathbf{y}\mathbf{y}^T\} - \mathbb{E}\{\mathbf{y}\mathbf{y}^T\})\mathbf{H} / \mathbf{H}^T\mathbb{E}\{\mathbf{y}\mathbf{y}^T\}\mathbf{H}$ maximally sparse.

In Section IV, we have shown that the independent source model, which is the basis for ICA, is also a suitable MaxViT model under a not too restrictive condition (i.e., the condition is generally satisfied by the sources because of their independence). Within this perspective, MaxViT (and thus also BSSR at order $2p = 2$) may be seen as a direct competitor to cICA [10]. Where the methods of cICA are generally based on the formation of an augmented Lagrangian in the framework of constrained programming using iterative updating methods and Newton iterations, the contrast function in MaxViT offers a closed form estimator for the extraction filter. Contrary to the family of cICA algorithms, we can now guarantee a global optimiser in the low-noise conditions. Moreover, in the noiseless case and for independent generating sources, MaxViT will provide a filter estimate from which we can obtain the independent source, of course under the condition that we can approximate the conditional set \mathcal{C}_{s_j} sufficiently well. A simple indicator function $\mathcal{I}_{s_j, \mathcal{C}}(k) = 1$ if $k \in \mathcal{C}_{s_j}$ and 0 otherwise, can now be seen as a simple binary reference signal. It should be investigated whether an appropriate weighing of $\mathcal{I}_{s_j, \mathcal{C}}(k)$ (e.g., as a function of the confidence one has in the relation $k \in \mathcal{C}_{s_j}$), getting a waveform rather than a binary reference, would yield better results.

Another drawback of the cICA based methods, is the constraint on the solution space to the correlation hyper cone around the reference signal, which is defined as $\mathcal{H}_c = \{x : |\mathbb{E}\{xr\}| / (\mathbb{E}\{x^2\}^{1/2}\mathbb{E}\{r^2\}^{1/2}) \leq c\}$. This requires a constant c to be set, separating the search space in an admissible and inadmissible correlation half space (the hypercone \mathcal{H} acting as the separating surface). The choice of the value attributed to the constant has a direct consequence on the convergence of the algorithms and is generally not intuitive [10]. As explained in Section III, the only constant in MaxViT that has to be set, is C . Fortunately, from Table I we see that its value is not critical, at least for large sample populations. In practical situations, where only a small population sample is available, it should neither be taken too large, nor too small, respectively because the conditional correlation would be calculated on a too small

sample set or because the condition in Eq. (5) would no longer be met. Even more, it should be remarked that in practice we can use an arbitrary function $f(s_j)$ in the conditional instead of a simple threshold on its absolute amplitude. Its interest may be to lower $|\rho|/\delta$ on the set \mathcal{C}_{s_j} . However, this is out of the scope of this paper and should be further investigated.

VII. CONCLUSION

We have shown that the maximising the likelihood criterion conditioned on a presence indicator gives rise to a contrast function for the extraction of a source of interest. The filter corresponding to the optimum of the contrast function can be found algebraically, provided that the conditional second moment can be estimated from the observations. The MaxViT estimator has interesting properties, such as robustness to noise or perturbations of the conditional set, making it a valuable alternative to constrained ICA algorithms.

APPENDIX

A. Proof of Proposition 1

Since the log-likelihood is either 0 or $(-\infty)$, and $\mathcal{L}(x) = 0$ holds if and only if we have $p_x(\bar{\mathcal{B}} | \mathcal{I}_{s_j}) = 0$, it remains to prove that the latter condition implies the equality $x = \lambda_j s_j$.

Proof: Suppose that we have $x \neq \lambda_j s_j$, and thus $x = \mathbf{g}^T \mathbf{s} = g_j s_j + \sum_{i \neq j} g_i s_i = g_j s_j + \check{s}$, where at least one g_i has a non-zero value and for which $p_x(\bar{\mathcal{B}} | \mathcal{I}_{s_j}) = 0$.

Since s_j is independently distributed with respect to \check{s} and thus with respect to all $s_i (i \neq j)$, we have that s_j is independently distributed with respect to \check{s} . As a consequence, the distribution $p_x(u | \mathcal{I}_{s_j})$ can be written as the convolution of the distributions $p_{s_j}(u | \mathcal{I}_{s_j})$ and $p_{\check{s}}(u)$, or

$$p_x(u | \mathcal{I}_{s_j}) = \int_{\mathbb{R}} p_{s_j}(\tau - u | \mathcal{I}_{s_j}) p_{\check{s}}(\tau) d\tau . \quad (20)$$

A necessary condition for x to yield $\mathcal{L}(x) = 0$, is that $p_x(u | \mathcal{I}_{s_j}) = 0, \forall u \in \bar{\mathcal{B}}$. However, if $\exists \varepsilon$ with non-zero Lebesgue measure for which the support set \mathcal{S} of $p_{\check{s}}$ has a measure $|\mathcal{S}| \geq \varepsilon$ and for which $p_{s_j}(C \leq |u| \leq C + |\varepsilon| | \mathcal{I}_{s_j}) > 0$, then, by Eq. (20), $\exists u \in \bar{\mathcal{B}} : p_x(u | \mathcal{I}_{s_j}) > 0$. As a consequence, our initial supposition was wrong and we must have $g_i = 0, \forall i \neq j$, i.e. $\mathcal{L}(x) = 0 \Rightarrow x = \lambda_j s_j$. ■

B. The objective function of Eq. (10) is a contrast for the extraction of s_j .

To proof that Ψ_x is a contrast under the conditions (C1)-(C3) from Section III-D, we need to show that it has the properties (P1')-(P3') from Section II-B.

Proof: The indeterminacy of the source scaling has been taken care of by the denominator in Eq. (10), and thus (P2') holds.

Furthermore we have

$$\frac{\mathbf{h}^T \Phi_{\mathbf{y}}^{s_j} \mathbf{h}}{\mathbf{h}^T \Phi_{\mathbf{y}} \mathbf{h}} = \frac{\mathbf{g}^T \Phi_{\mathbf{s}}^{s_j} \mathbf{g}}{\mathbf{g}^T \Phi_{\mathbf{s}} \mathbf{g}} = \frac{\sum_i |g_i|^2 \Phi_{s_i}^{s_j}}{\sum_i |g_i|^2} ,$$

since our sources are uncorrelated, both conditionally and unconditionally. Splitting up the sum in the different contributions gives us (up to a multiplicative positive constant)

$$|g_{jj}|^2 \Phi_{s_j}^{s_j} + \sum_{i \neq j} |g_i|^2 \Phi_{s_i}^{s_j} + \sum_{i \neq j} |g_i|^2 \Phi_{s_j}^{s_i} - \sum_{i \neq j} |g_i|^2 \Phi_{s_j}^{s_j}$$

which can be rewritten as

$$\Phi_{s_j}^{s_j} + \sum_{i \neq j} |g_i|^2 (\Phi_{s_i}^{s_j} - \Phi_{s_j}^{s_i}) \leq \Phi_{s_j}^{s_j} ,$$

where the inequality follows from $(\Phi_{s_i}^{s_j} - \Phi_{s_j}^{s_i}) < 0, \forall i \neq j$. This proves the domination.

We also have

$$\Phi_x^{s_j} = \Phi_{s_j}^{s_j} \Leftrightarrow \sum_{i \geq 2} |g_i|^2 (\Phi_{s_i}^{s_j} - \Phi_{s_1}^{s_j}) = 0 .$$

Now, since $(\Phi_{s_i}^{s_j} - \Phi_{s_1}^{s_j}) < 0, \forall i \neq j$, we have the above equality if and only if $|g_i|^2 = 0, \forall i \geq 2$. And thus

$$\frac{\mathbf{g}^H \Phi_{\mathbf{s}}^{s_j} \mathbf{g}}{\mathbf{g}^H \Phi_{\mathbf{s}}^R \mathbf{g}} = \frac{\Phi_{s_1}^{s_j}}{\Phi_{s_1}}$$

This proves the discrimination and thus, together with the domination, (P1') and (P3') are fulfilled.

Since any objective function fulfilling (P1')-(P3') is a contrast function for source extraction, our function $\Psi(x)$ in Eq. (10) is a contrast under the conditions (C1)-(C3). ■

Note that this could further be extended to the case where the covariance $\mathbb{E}_{s_j} \{s_i s_k\} \neq 0, \forall i, k \neq j, i \neq k$, as long as $\mathbb{E}_{s_j} \{s_j s_i\} = 0, \forall i \neq j$. For the proof, define $\tilde{\mathbf{s}} = [s_1, s_2 \dots s_{j-1}, s_{j+1} \dots s_N]^T$. Now take the eigenvalue decomposition of $\Phi_{\tilde{\mathbf{s}}}^{s_j}$ as $\mathbf{V}^T \Phi_{\tilde{\mathbf{s}}}^{s_j} \mathbf{V} = \Delta$, where Δ is a diagonal matrix with the eigenvalues λ_i on its diagonal and extend \mathbf{V} to

$$\tilde{\mathbf{V}} = \begin{pmatrix} 1 & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & \mathbf{V} \end{pmatrix} ,$$

where $\mathbf{0}_{n-1}$ is a vector of zeros in \mathbb{R}^{n-1} . The proof then continues similarly as above but replacing $\Phi_{\tilde{\mathbf{s}}}^{s_j}$ by Δ and \mathbf{g} by $\tilde{\mathbf{V}} \mathbf{g}$. As a consequence, condition (C3) in Section III-D becomes $\Phi_{s_j}^{s_j} > \max \lambda_i$.

C. Algebraic Solution for the 2×2 Case

Suppose that \mathbf{y} has uncorrelated, unit-variance and zero mean entries, without loss of generalisation, since \mathbf{y} can always be rendered so through whitening. Since $\Phi_{\mathbf{y}} = \mathbf{I}_2$, the eigenvector that would separate our source as $x = \mathbf{e}_1^T \mathbf{y}$ is the dominant eigenvector of the covariance matrix $\Phi_{\mathbf{y}}^{s_j}$, which has a general symmetric form

$$\Phi_{\mathbf{y}}^{s_j} = \begin{bmatrix} a & b \\ b & c \end{bmatrix} . \quad (21)$$

The above matrix has eigenvalues

$$\lambda_{1,2} = \frac{a+c}{2} \pm \sqrt{\left(\frac{a-c}{2}\right)^2 + |b|^2} , \quad (22)$$

and thus, if $a \neq c$, has a largest eigenvector

$$\mathbf{e}_1 = \begin{bmatrix} \pm 1 \frac{1}{\sqrt{2}} \sqrt{1 + \frac{\xi}{\xi^2 + |b|^2}} \\ \pm 1 \frac{1}{\sqrt{2}} \sqrt{1 - \frac{\xi}{\xi^2 + |b|^2}} \end{bmatrix}, \quad (23)$$

with $\xi = \frac{a-c}{2}$.

D. Calculation of ISR

We induce the estimation bound in case the sources are not perfectly conditionally uncorrelated. Since we only consider unitary transformations $\mathbf{A} = \mathbf{Q}$ (for our \mathbf{y} is or has been rendered spatially *white*), we know that the eigenvalues of $\Phi_{\mathbf{s}}^{s_j}$ and $\Phi_{\mathbf{y}}^{s_j}$ are equal. Moreover, the i -th eigenvector \mathbf{q}_i of $\Phi_{\mathbf{y}}^{s_j}$ equals $\mathbf{Q}\mathbf{e}_i$, where \mathbf{e}_i is the i -th eigenvector of $\Phi_{\mathbf{s}}^{s_j}$ (see also the equivariance property [20]). As a consequence, we only need to consider the simpler case where $\mathbf{Q} = \mathbf{I}_m$, without loss of generality.

Limiting \mathbf{s} to belong to \mathbb{R}^2 , the matrix $\Phi_{\mathbf{s}}^{s_j}$ takes the form

$$\Phi_{\mathbf{s}}^{S_c(s_1)} = \begin{bmatrix} \Phi_{s_1}^{s_j} & \mathbb{E}_{s_j}\{s_1 s_2\} \\ \mathbb{E}_{s_j}\{s_1 s_2\} & \Phi_{s_2}^{s_j} \end{bmatrix}. \quad (24)$$

From Eq. (23), one can explicitly calculate the separation filter \mathbf{h} associated to $\Phi_{\mathbf{s}}^{s_1}$ by the above Eq. (23). As such we obtain for the ISR ($|g_2|^2/|g_1|^2 = |h_2|^2/|h_1|^2$):

$$\text{ISR} = \frac{\sqrt{\delta^2 + |\rho|^2} - \delta^2}{\sqrt{\delta^2 + |\rho|^2} + \delta^2}, \quad (25)$$

with $\delta = \frac{\Phi_{s_1}^{s_1} - \Phi_{s_2}^{s_1}}{2}$ and $\rho = \mathbb{E}_{s_j}\{s_1 s_2\}$.

E. Proof of the inequality $\mathbb{E}_{s_j}\{s_j^2\} \geq \mathbb{E}\{g(s_j^2)\}$

Proof: To prove the inequality, we proof the more general form $\mathbb{E}_{s_j}\{g(s_j)\} \geq \mathbb{E}\{g(s_j)\}$ for any positive valued function g . We have

$$\mathbb{E}_{s_j}\{g(s_j)\} = \int p_{s_j}(u | \mathcal{I}_{s_j}) g(u) du \quad (26)$$

$$= \frac{\int_{\mathcal{B}} p_{s_j}(u) g(u) du}{P_{s_j}(\mathcal{B})} \quad (27)$$

$$= \frac{\int p_{s_j}(u) g(u) du - \int_{\mathcal{B}} p_{s_j}(u) g(u) du}{P_{s_j}(\mathcal{B})} \quad (28)$$

$$> \int p_{s_j}(u) g(u) du = \mathbb{E}\{g(s_j)\}, \quad (29)$$

These results hold if we impose the condition of Eq. (5), since we have from Hölders inequality that $\int_{\mathcal{B}} p_{s_j}(u) du \int_{\mathcal{B}} g(u) du \geq \int_{\mathcal{B}} p_{s_j}(u) g(u) du$ and thus $\int_{\mathcal{B}} p_{s_j}(u) du \rightarrow 0 \Rightarrow \int_{\mathcal{B}} p_{s_j}(u) g(u) du \rightarrow 0$ for all positive valued functions g . As a consequence, we have $\mathbb{E}_{s_j}\{g(s_j)\} \geq \mathbb{E}\{g(s_j)\}$ with equality if and only if $P_{s_j}(\mathcal{B}) = 1$, i.e., $C = 0$. Since u^2 is a non-negative valued function and $C > 0$, we have $\mathbb{E}_{s_j}\{s_j^2\} > \mathbb{E}\{s_j^2\}$. ■

REFERENCES

- [1] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.
- [2] C. Jutten and J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.
- [3] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, pp. 287–314, 1994.
- [4] L. Tong, R.-W. Liu, V. C. Soon, and Y.-F. Huang, "Indeterminacy and identifiability of blind identification," *IEEE Trans on Circuits and Systems*, vol. 38, no. 5, pp. 499–509, 1991.
- [5] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Processing*, vol. 45, pp. 59–83, 1995.
- [6] A. Hyvarinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
- [7] S. A. Cruces-Alvarez, A. Cichocki, and S.-i. Amari, "From blind signal extraction to blind instantaneous signal separation: Criteria, algorithms and stability," *IEEE Transactions on Neural Networks*, vol. 15, no. 4, pp. 859–873, 2004.
- [8] A. Cichocki, R. Thawonmas, and S. ichi Amari, "Sequential blind signal extraction in order specified by stochastic properties," *Electronics Letters*, vol. 33, no. 1, pp. 64–65, 1997.
- [9] W. Lu and J. C. Rajapakse, "ICA with reference," in *Proc. Int. Conf. on ICA and BSS*, 2001, pp. 120 – 125.
- [10] —, "Approach and applications of constrained ica," *IEEE Trans on Neural Networks*, vol. 16, no. 1, pp. 203–212, 2005.
- [11] M. Sato, Y. Kimura, S. Chida, T. Ito, N. Katayama, K. Okamura, and M. Nakao, "A novel extraction method of fetal electrocardiogram from the composite abdominal signal," *IEEE Trans on Biom Eng*, vol. 54, no. 1, pp. 49–58, 2007.
- [12] A. Adib, E. Moreau, and D. Aboutajdine, "Source separation contrasts using a reference signal," *IEEE Signal Processing Letters*, vol. 11, no. 3, pp. 312–315, 2004.
- [13] M. Castella, S. Rhioui, E. Moreau, and J. C. Pesquet, "Quadratic higher order criteria for iterative blind separation of a mimo convolutive mixture of sources," *IEEE Trans. Signal Process.*, vol. 55, no. 1, pp. 218–232, Jan. 2007.
- [14] D. Donoho, *On Minimum Entropy Deconvolution*. Academic Press, 1981, ch. On Minimum Entropy Deconvolution, pp. 565–608.
- [15] E. Moreau and P. Comon, *Séparation de sources*. Hermès-Lavoisier, 2007, vol. 1, ch. Fonctions de Contraste, pp. 75–115, in french.
- [16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.
- [17] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley Interscience, 2001.
- [18] A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, 2005th ed. Wiley, 2002.
- [19] S. Roberts and R. Everson, Eds., *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.
- [20] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, 1996.
- [21] C. J. James and O. Gibson, "ICA with a reference: extracting desired electromagnetic brain signals," *Medical Applications of Signal Processing*, 2002.
- [22] C. J. James and O. J. Gibson, "Temporally constrained ICA: an application to artifact rejection in electromagnetic brain signal analysis," *IEEE Trans on Biomed Eng*, vol. 30, no. 9, pp. 1108–1115, 2003.
- [23] T. Netabayashi, Y. Kimura, S. Chida, T. Ito, K. Ohwada, N. Katayama, K. Okamura, and M. Nakao, "Robustness of the blind source separation with reference against uncertainties of the reference signals," in *30th Annual International IEEE EMBS Conference*, vol. 30, Vancouver, British Columbia, Canada, 2008, pp. 1875–1878.
- [24] R. Phlypo, V. Zarzoso, P. Comon, and I. Lemahieu, "Cumulant matching for independent source extraction," in *30th Annual International IEEE EMBS Conference*, Vancouver, British Columbia, Canada, 2008, pp. 3340–3343.
- [25] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
- [26] R. Gribonval, "Piecewise linear source separation," in *Proc. SPIE'03, ser. Wavelets: Applications in Signal and Image Processing*, vol. 5207, San Diego, California, USA, 2003.
- [27] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.