# Orthogonally-Constrained Extraction of Independent Non-Gaussian Component from Non-Gaussian Background Without ICA

Zbyněk Koldovský[1(✉)], Petr Tichavský[2], and Nobutaka Ono[3]

[1] Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
zbynek.koldovsky@tul.cz

[2] The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod vodárenskou věží 4, P.O. Box 18, 182 08 Praha 8, Czech Republic
tichavsk@utia.cas.cz

[3] Faculty of System Design, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan
onono@tmu.ac.jp

**Abstract.** We propose a new algorithm for Independent Component Extraction that extracts one non-Gaussian component and is capable to exploit the non-Gaussianity of background signals without decomposing them into independent components. The algorithm is suitable for situations when the signal to be extracted is determined through initialization; it shows an extra stable convergence when the target component is dominant. In simulations, the proposed method is compared with Natural Gradient and One-unit FastICA, and it yields improved results in terms of the Signal-to-Interference ratio and the number of successful extractions.

**Keywords:** Independent Component Analysis
Blind source separation · Non-Gaussian distribution · Score function
Independent Vector Analysis

## 1 Introduction

The Blind Source Extraction (BSE) problem where the goal is to extract one particular component from a linear mixture

$$\mathbf{x} = \mathbf{A}\mathbf{u}, \tag{1}$$

has been a live topic for decades, also before the birth of Independent Component Analysis (ICA) [3,4,7]. In the mixture, $\mathbf{u}$ and $\mathbf{x}$ are $d \times 1$ vectors, respectively,

This work was supported by The Czech Science Foundation through Project No. 17-00902S and partially supported by JSPS KAKENHI Grant Number 16H01735.


Y. Deville et al. (Eds.): LVA/ICA 2018, LNCS 10891, pp. 161–170, 2018.
https://doi.org/10.1007/978-3-319-93764-9_16

of $d$ original and mixed signals, and $\mathbf{A}$ is a $d \times d$ non-singular mixing matrix. The components of $\mathbf{u}$ are assumed to be *mutually independent*. Let, without any loss of generality, the desired component be $u_1$, which will be referred to as SOI (the source of interest); the other signals will be briefly called *background*.

By information theory, it is possible to extract an independent component through finding a direction having minimum entropy (maximum non-Gaussianity). However, methods extracting one non-Gaussian independent component in this way (from here referred to as "one-unit" methods) are known to have a limited asymptotic accuracy compared to methods performing the whole ICA decomposition of (1). Performance analyses of several one-unit methods showed that they perform as if background components were all Gaussian [5,12,14].

Specifically, let $\mathbf{W}$ be an unbiased estimate of $\mathbf{A}^{-1}$ (a de-mixing matrix) up to the order and scales of its rows, and $\mathbf{G} = \mathbf{WA} \approx \mathbf{P}\boldsymbol{\Lambda}$, where $\mathbf{P}$ and $\boldsymbol{\Lambda}$ is, respectively, a permutation and a diagonal matrix. The Cramér-Rao bound (CRLB) for ICA says that [14,15]

$$\mathrm{E}[G_{ij}^2] \geq \frac{1}{N} \frac{\kappa_j}{\kappa_i \kappa_j - 1}, \qquad i \neq j, \tag{2}$$

where $\mathrm{E}[\cdot]$ stands for the expectation operator, $N$ is the number of samples of $\mathbf{x}$ (assuming identically and independently distributed samples), and $\kappa_i = \mathrm{E}[\psi_i^2]$ where $\psi_i(x) = -\partial/\partial x \, \log p_i(x)$, which is the score function of $p_i$ where $p_i$ is the pdf of the $i$th original signal $u_i$.

For normalized variables with unit variance it holds that $\kappa_i \geq 1$ where $\kappa_i = 1$ if and only if the $i$th pdf is Gaussian. Let $\mathbf{w}$ be the first row of $\mathbf{W}$ corresponding to the extracted SOI, and let $\mathbf{u}$ have all unit variance. The asymptotic accuracy (for $N \to +\infty$) of one-unit methods (when the true score function is used in the algorithm's contrast function) was shown to be characterized by [5,12,14]

$$\mathrm{E}[g_j^2] \approx \frac{1}{N} \frac{1}{\kappa_1 - 1}, \qquad j \neq 1, \tag{3}$$

where $\mathbf{g} = \mathbf{w}^T\mathbf{A}$. The right-hand side coincides with the CRLB in (2) for $i = 1$ when $\kappa_j = 1$ for $j = 2, \ldots, d$, which is the case when $u_2, \ldots, u_d$ are Gaussian (for which case the CRLB (2) formally does not exist unless $d = 2$).

Recently, we have revised the BSE problem through Independent Component Extraction (ICE) [10,11]. Here, the mixing model (1) is re-parameterized for the extraction of the SOI in the way that the rest of the mixture is not object of any particular decomposition, as compared to ICA. In the statistical model, $s$ is assumed to be non-Gaussian while the other components are assumed to be Gaussian. Under these conditions, the CRLB for ICE has been confirmed to correspond to the right-hand side of (3); see [8]. In [10], orthogonally-constrained gradient learning algorithms for ICE have been proposed based on the maximum likelihood principle.[1] An appealing property of these algorithms resides in their

[1] A particular variant of these algorithms ($\mathrm{OGICE}_{\mathbf{w}}$) coincides with a method proposed earlier by Pham in [12], which was derived based on a simplified form of mutual information that is valid for Gaussian background.

ability to keep converging to the desired source, e.g., to a dominant SOI. Using methods that guarantee the extraction of the SOI with a high probability, the complete ICA decomposition and the subsequent component selection due to the random order can be avoided, which brings significant computational savings.

In this paper, our goal is to overcome the accuracy limitation given by (3). We derive a new gradient ICE algorithm using the maximum likelihood approach. The method takes into account possible non-Gaussianity of background. For simplicity, real-valued mixing scenario and signals will be considered, although a complex-valued extension is possible.

The rest of this paper is organized as follows. The ICE mixing model and the statistical model of signals are described in Sect. 2. In Sect. 3, the novel algorithm is proposed and described in details. Section 4 is devoted to simulations and comparisons, and Sect. 5 concludes the paper.

Notation: Plain letters denote scalars; bold letters denote vectors; bold capital letters denote matrices. The Matlab convention 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 $\mathbf{A}$. Symbolic scalar and vector random variables will be denoted by lower case letters, e.g. $s$ and $\mathbf{x}$, $\mathbf{z}$, while the quantities collecting their $N$ samples will be denoted by bold (capital) letters, e.g. $\mathbf{s}$ (a row vector $1 \times N$) and $\mathbf{X}$, $\mathbf{Z}$. Estimated values of signals will be denoted by hat, e.g., $\widehat{\mathbf{s}}$, $\widehat{\mathbf{Z}}$.

## 2 Problem Formulation

### 2.1 Algebraic Mixing Model

Let the SOI be $s = u_1$ and $\mathbf{a}$ be the first column of $\mathbf{A}$, so $\mathbf{A}$ can be partitioned as $\mathbf{A} = [\mathbf{a}, \mathbf{A}_2]$. Then, $\mathbf{x}$ can be written as

$$\mathbf{x} = \mathbf{a}s + \mathbf{y}, \tag{4}$$

where $\mathbf{y} = \mathbf{A}_2\mathbf{u}_2$ and $\mathbf{u}_2 = [u_2, \ldots, u_d]^T$. The fact that $\mathbf{y} = \mathbf{A}_2\mathbf{u}_2$ means that the mixture consists of the same number of sources as that of input channels.

Let the new parameterization of the mixing matrix and of its inverse matrix be denoted by $\mathbf{A}_{\mathrm{ICE}}$ and $\mathbf{W}_{\mathrm{ICE}}$, respectively. In ICE, the identification of $\mathbf{A}_2$ or the decomposition of $\mathbf{y}$ into independent signals $\mathbf{u}_2$ is *not* the goal. Therefore, we assume that $\mathbf{A}_{\mathrm{ICE}} = [\mathbf{a}, \mathbf{Q}]$ where $\mathbf{Q}$ is, for now, arbitrary with full column-rank. Then, (4) can be written as

$$\mathbf{x} = \mathbf{A}_{\mathrm{ICE}}\mathbf{v}, \tag{5}$$

where $\mathbf{v} = [s; \mathbf{z}]$, and $\mathbf{y} = \mathbf{Q}\mathbf{z}$. Hence, the subspace spanned by $\mathbf{z}$ is the same as that of $\mathbf{u}_2$. To complete the mixing matrix definition, we look at the inverse matrix $\mathbf{W}_{\mathrm{ICE}} = \mathbf{A}_{\mathrm{ICE}}^{-1}$.

Let $\mathbf{a}$ and $\mathbf{W}_{\mathrm{ICE}}$ be partitioned, respectively, as $\mathbf{a} = [\gamma; \mathbf{g}]$ and $\mathbf{W}_{\mathrm{ICE}} = [\mathbf{w}^T; \mathbf{B}]$. $\mathbf{B}$ is required to be orthogonal to $\mathbf{a}$, i.e. $\mathbf{B}\mathbf{a} = \mathbf{0}$, which ensures that $\mathbf{B}\mathbf{x}$ do not contain any contribution of $s$. A useful selection is $\mathbf{B} = [\mathbf{g}, -\gamma\mathbf{I}_{d-1}]$

where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. Let $\mathbf{w}$ be partitioned as $\mathbf{w} = [\beta; \mathbf{h}]$. Then,

$$\mathbf{W}_{\mathrm{ICE}} = \begin{pmatrix} \mathbf{w}^T \\ \mathbf{B} \end{pmatrix} = \begin{pmatrix} \beta & \mathbf{h}^T \\ \mathbf{g} & -\gamma \mathbf{I}_{d-1} \end{pmatrix}, \tag{6}$$

and from $\mathbf{A}_{\mathrm{ICE}} \cdot \mathbf{W}_{\mathrm{ICE}} = \mathbf{I}_d$ it follows that

$$\mathbf{A}_{\mathrm{ICE}} = [\mathbf{a},\, \mathbf{Q}] = \begin{pmatrix} \gamma & \mathbf{h}^T \\ \mathbf{g} & \left(\mathbf{g}\mathbf{h}^T - \mathbf{I}_{d-1}\right)\gamma^{-1} \end{pmatrix}, \tag{7}$$

where $\beta$ and $\gamma$ are linked through

$$\beta\gamma = 1 - \mathbf{h}^T\mathbf{g}. \tag{8}$$

The latter equation can be also written in the form $\mathbf{w}^T\mathbf{a} = 1$, which corresponds to the *distortionless response* constraint [16]. The role of $\mathbf{a}$, as follows from (4), is the *mixing vector* related to $s$, while $\mathbf{w}$ is the *separating vector* as $s = \mathbf{w}^T\mathbf{x}$. For the background signal $\mathbf{z}$, it holds that $\mathbf{z} = \mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{A}_2\mathbf{u}_2$.

Similarly to the indeterminacies in ICA, the scales of $s$ and of $\mathbf{a}$ are ambiguous in the sense that they can be replaced, respectively, by $\alpha s$ and $\alpha^{-1}\mathbf{a}$ where $\alpha \neq 0$. The scaling ambiguity can be avoided by fixing $\beta$ or $\gamma$. Next, the role of $s = u_1$ can be interchanged with $u_i$, for any $i = 2, \ldots, d$. This is the permutation problem [13].

In this paper, we assume that an initial guess of $\mathbf{a}$ or of $\mathbf{w}$ is given, which determines the SOI. The initial value is typically deviated by an estimation error, which increases the probability that the given algorithm finally extracts a different source than the SOI. In experiments (Sect. 4), we therefore conduct a sensitivity analysis, which compares the size of the attraction area of different BSE algorithms.

### 2.2 Statistical Model

The main principle of ICE is the same as that of ICA. We take the assumption that $s$ and $\mathbf{z}$ are *independent*, so the goal is to find $\mathbf{a}$ and $\mathbf{w}$ such that $\mathbf{w}^T\mathbf{x}$ and $\mathbf{B}\mathbf{x}$ are independent (or as independent as possible).

Let the pdf of $s$ and of $\mathbf{z}$ be, respectively, $p_s(s)$ and $p_{\mathbf{z}}(\mathbf{z})$. The joint pdf of the mixed signals $\mathbf{x} = \mathbf{A}_{\mathrm{ICE}}\mathbf{v}$ is

$$p_{\mathbf{x}}(\mathbf{x}) = p_s(\mathbf{w}^T\mathbf{x}) \cdot p_{\mathbf{z}}(\mathbf{B}\mathbf{x}) \cdot |\det \mathbf{W}_{\mathrm{ICE}}| \tag{9}$$

where it can be shown that

$$\det \mathbf{W}_{\mathrm{ICE}} = (-1)^{d-1}\gamma^{d-2} = (-1)^{d-1}\beta^{-(d-2)}(1 - \mathbf{h}^T\mathbf{g})^{d-2}. \tag{10}$$

Since the background signals remain unmixed after ICE (up to special cases such as $d = 2$), we proposed in [10,11] to model the unknown $p_{\mathbf{z}}$ as Gaussian with zero mean and covariance $\mathbf{C}_{\mathbf{z}}$. In this paper, we generalize the background model to arbitrary (non-)Gaussian pdf. Thus, the unknown densities $p_s(s)$ and

$p_{\mathbf{z}}(\mathbf{z})$ are replaced, respectively, by model densities $f(s)$ and $q(\mathbf{z})$. The quasi-loglikelihood function for $N$ i.i.d. signal samples, according to (9), takes the form

$$\mathcal{L}(\mathbf{a},\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\left\{\log f(\mathbf{w}^T\mathbf{x}(n)) + \log q(\mathbf{B}\mathbf{x}(n))\right\} + (d-2)\log|\gamma|. \tag{11}$$

**Orthogonal Constraint.** The first term on the right-hand side of (11) depends purely on $\mathbf{w}$, while the second and the third terms depend purely on $\mathbf{a}$. The only link between $\mathbf{a}$ and $\mathbf{w}$ thus resides in (8). Therefore, the likelihood function can have spurious maxima where $\mathbf{a}$ and $\mathbf{w}$ do not correspond to the same source.

To make the interconnection between $\mathbf{a}$ and $\mathbf{w}$ tighter, the orthogonal constraint (OG) can be imposed [2]. Let $\mathbf{W}_{\mathrm{ICE}}$ be a current ICE de-mixing matrix estimate having the structure of (6), and $\widehat{\mathbf{V}} = [\widehat{\mathbf{s}}; \widehat{\mathbf{Z}}] = \mathbf{W}_{\mathrm{ICE}}\mathbf{X}$ be the estimated matrix of de-mixed signal samples. The OG reads

$$\frac{1}{N}\widehat{\mathbf{s}}\cdot\widehat{\mathbf{Z}}^T = \frac{1}{N}\mathbf{w}^T\mathbf{X}\mathbf{X}^T\mathbf{B}^T = \mathbf{w}^T\widehat{\mathbf{C}_{\mathbf{x}}}\mathbf{B}^T = \mathbf{0}, \tag{12}$$

where $\widehat{\mathbf{C}_{\mathbf{x}}} = \mathbf{X}\mathbf{X}^T/N$ is the sample-based estimate of $\mathbf{C}_{\mathbf{x}} = \mathrm{E}[\mathbf{x}\mathbf{x}^T]$. The reader can verify that the OG together with (8) introduce the following links between $\mathbf{a}$ and $\mathbf{w}$:

$$\mathbf{a} = \frac{\widehat{\mathbf{C}_{\mathbf{x}}}\mathbf{w}}{\mathbf{w}^T\widehat{\mathbf{C}_{\mathbf{x}}}\mathbf{w}}, \tag{13a}$$

$$\mathbf{w} = \frac{\widehat{\mathbf{C}_{\mathbf{x}}}^{-1}\mathbf{a}}{\mathbf{a}^T\widehat{\mathbf{C}_{\mathbf{x}}}^{-1}\mathbf{a}}. \tag{13b}$$

In this paper, we will consider the former coupling, that is, $\mathbf{w}$ will be the free variable while $\mathbf{a}$ will be treated as dependent.

## 3 Gradient-Based Algorithm

### 3.1 Gradient of the Contrast Function

The gradient of $\mathcal{L}$ with respect to $\mathbf{w}$ under the coupling (13a), is

$$\left.\frac{\partial\mathcal{L}}{\partial\mathbf{w}}\right|_{\text{w.r.t. (13a)}} = -\frac{1}{N}\mathbf{X}\widehat{\boldsymbol{\phi}}^T + \frac{1}{N}\frac{\widehat{\mathbf{C}}_{\mathbf{x}}}{\mathbf{w}^T\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{w}}\begin{pmatrix}\mathtt{tr}(\mathbf{E}\mathbf{X}\widehat{\boldsymbol{\Psi}}) + (d-2)N\gamma^{-1}\\ -\widehat{\boldsymbol{\Psi}}\mathbf{X}^T\mathbf{e}_1\end{pmatrix} + 2\mathbf{a}\left(\frac{1}{N}\mathtt{tr}(\widehat{\boldsymbol{\Psi}}^T\widehat{\mathbf{Z}}) - (d-2)\right), \tag{14}$$

where $\mathtt{tr}(\cdot)$ denotes the trace, $\mathbf{E} = [\mathbf{0}, \quad \mathbf{I}_{d-1}]$, $\mathbf{e}_1$ denotes the first column of $\mathbf{I}_d$, $\widehat{\boldsymbol{\phi}} = \phi(\widehat{\mathbf{s}})$, and $\widehat{\boldsymbol{\Psi}} = \psi(\widehat{\mathbf{Z}})$, where

$$\phi(\xi) = -\frac{\partial\log f(\xi)}{\partial\xi} \quad \text{and} \quad \psi_i(\mathbf{z}) = -\frac{\partial\log q(\mathbf{z})}{\partial z_i}, \quad \psi(\mathbf{z}) = [\psi_1(\mathbf{z}),\ldots,\psi_{d-1}(\mathbf{z})]^T, \tag{15}$$

are the score function of the model pdfs $f(\cdot)$ and $q(\cdot)$, respectively, which are applied element/column-wise in case of the vector/matrix argument. We skip details of the lengthy computation of (14) here due to the lack of space.

By exploring this gradient when $N \to +\infty$ and when $\mathbf{w}$ is the ideal separating vector, that is, when $\mathbf{w}^T\mathbf{x} = s$ and $\mathbf{B}\mathbf{x} = \mathbf{z}$, an important fact can be shown: The ideal separating vector is a stationary point of the contrast function (the gradient is zero) only if $\phi$ and $\psi$ satisfy $\mathrm{E}[s\phi(s)] = 1$ and $\mathrm{E}[\mathbf{z}\psi(\mathbf{z})^T] = \mathbf{I}_{d-1}$, respectively. Both conditions are automatically satisfied when $\phi$ and $\psi$ are the true score functions of the respective variables. However, since these are not known in the blind scenario, we introduce the following normalizing conditions: For any estimates of $\mathbf{a}$ and $\mathbf{w}$, let

$$\widehat{\mathbf{s}}\widehat{\boldsymbol{\phi}}^T = N \qquad \text{and} \qquad \widehat{\mathbf{Z}}\widehat{\boldsymbol{\Psi}}^T = N\mathbf{I}_{d-1}. \tag{16}$$

With these conditions and after few computations, (14) simplifies to

$$\left.\frac{\partial \mathcal{L}}{\partial \mathbf{w}}\right|_{\text{w.r.t. (13a)}} = \mathbf{a} - \frac{1}{N}\mathbf{X}\widehat{\boldsymbol{\phi}}^T + \frac{1}{\mathbf{w}^T\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{w}}\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{B}^T\mathbf{p}, \tag{17}$$

where $\mathbf{p} = \widehat{\boldsymbol{\Psi}}\widehat{\mathbf{s}}^T/N$.

A practical way to select $\phi$ and $\psi$ meeting the conditions in (16) is by taking some appropriate prototype functions $\phi_1$ and $\psi_1$ instead. Then, the normalization can be done through defining

$$\widehat{\boldsymbol{\phi}} = N(\widehat{\mathbf{s}}\widehat{\boldsymbol{\phi}}_1^T)^{-1}\widehat{\boldsymbol{\phi}}_1 \qquad \text{and} \qquad \widehat{\boldsymbol{\Psi}} = \mathbf{R}^{-1}\widehat{\boldsymbol{\Psi}}_1, \tag{18}$$

where $\mathbf{R} = \widehat{\mathbf{Z}}\widehat{\boldsymbol{\Psi}}_1^T/N$, $\widehat{\boldsymbol{\phi}}_1 = \phi_1(\widehat{\mathbf{s}})$, and $\widehat{\boldsymbol{\Psi}}_1 = \psi_1(\widehat{\mathbf{Z}})$.

A special case that is worth to mention at this point is when the background signals $\mathbf{z}$ are Gaussian, i.e., $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{z}})$. The covariance $\mathbf{C}_{\mathbf{z}}$ is an unknown nuisance parameter, which must be replaced by the sample-based covariance of $\widehat{\mathbf{Z}}$, that is, by $\widehat{\mathbf{C}}_{\mathbf{z}} = \widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^T/N$. It means that the model density $q(\cdot)$ corresponds to $\mathcal{N}(\mathbf{0}, \widehat{\mathbf{C}}_{\mathbf{z}})$, whose score function is $\psi(\mathbf{z}) = \widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}$. Then, $\widehat{\boldsymbol{\Psi}} = \widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\widehat{\mathbf{Z}}$, $\mathbf{R} = \mathbf{I}_{d-1}$, and $\mathbf{p} = \widehat{\boldsymbol{\Psi}}\widehat{\mathbf{s}}^T/N = \widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\widehat{\mathbf{Z}}\widehat{\mathbf{s}}^T/N = \mathbf{0}$ due to the OG (12). Consequently, the third term on the right-hand side of (17) is zero, and the gradient simplifies to

$$\left.\frac{\partial \mathcal{L}}{\partial \mathbf{w}}\right|_{\text{w.r.t. (13a)}} = \mathbf{a} - \frac{1}{N}\mathbf{X}\widehat{\boldsymbol{\phi}}^T. \tag{19}$$

This result coincides with those derived in [10,12] under the Gaussian assumption.

The third term on the right-hand side of (17) can be seen as a correction term due to the non-Gaussianity of $\mathbf{z}$, as $\mathbf{p}$ consists of higher-order correlations between $\widehat{\mathbf{s}}$ and $\widehat{\mathbf{Z}}$, unless $\psi$ is purely linear.

### 3.2 Proposed Algorithm

We propose a gradient-based algorithm whose steps are described in Algorithm 1. In every step, the OG is imposed through (13a), the normalization steps given

by (18) are done, and the method updates $\mathbf{w}$ in the direction of the steepest ascent of $\mathcal{L}$. This is repeated until the norm of the gradient is smaller than `tol`; $\mu$ is the step length parameter; $\mathbf{w}_{\text{ini}}$ is the initial guess. We call this method OGICENGB.

**Algorithm 1.** OGICENGB: separating vector estimation based on orthogonally constrained gradient-ascent algorithm

**Input**: $\mathbf{X}$, $\mathbf{w}_{\text{ini}}$, $\mu$, `tol`, $\phi(\cdot)$, $\psi(\cdot)$
**Output**: $\mathbf{a}, \mathbf{w}$

$\widehat{\mathbf{C}}_{\mathbf{x}} = \mathbf{X}\mathbf{X}^T/N$;
$\mathbf{w} = \mathbf{w}_{\text{ini}}$;
**repeat**
  $\lambda_{\mathbf{w}} \leftarrow (\mathbf{w}^T\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{w})^{-1}$;
  $\mathbf{a} \leftarrow \lambda_{\mathbf{w}}\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{w}$; /* OG constraint (13a) */
  $\mathbf{B} = [\mathbf{a}_{2:d},\ -\mathbf{a}_1\mathbf{I}_{d-1}]$; /* by (6) */
  $\widehat{\mathbf{s}} \leftarrow \mathbf{w}^T\mathbf{X}$; /* current SOI estimate */
  $\widehat{\mathbf{Z}} \leftarrow \mathbf{B}\mathbf{X}$; /* current background estimate */
  $\nu \leftarrow \widehat{\mathbf{s}}\phi(\widehat{\mathbf{s}})^T/N$; /* normalizing constant from (18) */
  $\mathbf{T} \leftarrow \mathbf{X}\psi(\widehat{\mathbf{Z}})^T/N$; /* auxiliary matrix due to (18) */
  $\mathbf{p} = (\mathbf{B}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{w}$; /* by the definition of $\mathbf{p}$ */
  $\Delta \leftarrow \mathbf{a} - \nu^{-1}\mathbf{X}\phi(\widehat{\mathbf{s}})^T/N + \lambda_{\mathbf{w}}^{-1}\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{B}^T\mathbf{p}$; /* by (17) */
  $\mathbf{w} \leftarrow \mathbf{w} + \mu\Delta$; /* gradient ascent */
**until** $\|\Delta\| < \texttt{tol}$;

## 4 Simulations

We compare OGICENGB with its special variant OGICE (assuming Gaussian background) [10,12], with One-unit FastICA (FICA) from [6], and with the Natural Gradient algorithm (NG) for ICA [1]. In one trial, an instantaneous mixture of $d = 10$ signals is generated according to (1), and the SOI is extracted and evaluated in terms of Signal-to-Interference Ratio (SIR). The SOI $u_1$ as well as $u_2, \ldots, u_d$ are drawn from the Laplacean distribution. The scales of the components are selected so that $\text{SIR}_{\text{in}} = (d-1)\text{E}[|u_1|^2](\sum_{i=2}^{d}\text{E}[|u_i|^2])^{-1}$ is 10 dB. The elements of mixing matrices are drawn uniformly from $[1, 2]$, which ensures approximately equal SIR across all channels. The improvement of SIR ($\text{SIR}_{\text{imp}}$) is defined as the ratio between the average SIR on channels and the output SIR of the extracted source; the extraction is rated as *successful* if $\text{SIR}_{\text{imp}} > 0$ dB. The percentage of successful trials will be referred to as *success rate*.

The algorithms are initialized by $\mathbf{a}_{\text{ini}} = \mathbf{a} + \mathbf{e}_{\text{ini}}$, where $\mathbf{a}$ is the true mixing vector, and $\mathbf{e}_{\text{ini}}$ is a random vector with Gaussian entries such that $\|\mathbf{e}_{\text{ini}}\|^2 = \epsilon^2$. NG is initialized by a de-mixing matrix yielding background subspace that is

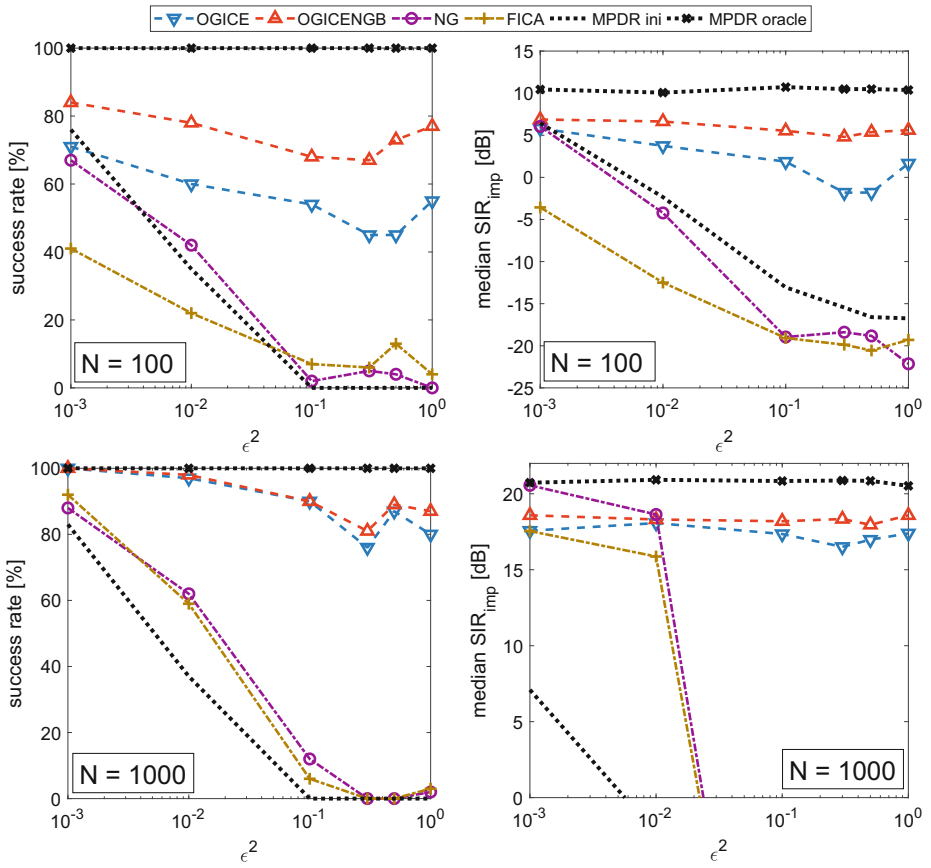


**Fig. 1.** Success rate and median SIR improvement as functions of $\epsilon^2$ achieved by the compared algorithms in 100 trials for $N = 100$ (row 1) and $N = 1000$ (row 2).

orthogonal to the initial SOI estimate[2]. To compare, the SOI estimates using (13b) with the true mixing vector (MPDR oracle) and with $\mathbf{a} = \mathbf{a}_{\text{ini}}$ (MPDR ini) are evaluated, also.

In algorithms, we choose $\phi(s) = \tanh(s)$, which is a smooth approximation of $\text{sign}(x)$ (the true score function for the Laplacean pdf). For choosing $\psi$ in OGICENGB, we adopt the idea from [9] for modeling dependent variables using the multivariate super-Gaussian distribution with covariance $\mathbf{C_z}$. Thus, the model pdf and the corresponding score function are, respectively,

$$q(\mathbf{z}) \propto \exp\left\{-\sqrt{\mathbf{z}^T\mathbf{C}_\mathbf{z}^{-1}\mathbf{z}}\right\} \quad \text{and} \quad \psi(\mathbf{z}) = \mathbf{C}_\mathbf{z}^{-1}\mathbf{z}/\sqrt{\mathbf{z}^T\mathbf{C}_\mathbf{z}^{-1}\mathbf{z}}. \tag{20}$$

[2] Note that the separated sources by NG are *not* reordered after the separation, because the BSE problem is assumed to be resolved correctly only if the SOI appears in the assumed output channel.

Based on this, our final choice of $\psi$ is $\psi(\mathbf{z}) = \widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}/\sqrt{\mathbf{z}^T\widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}}$ where $\widehat{\mathbf{C}}_{\mathbf{z}}$ is the sample-based estimate of $\mathbf{C}_{\mathbf{z}}$, namely, $\widehat{\mathbf{C}}_{\mathbf{z}} = \widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^T/N = \mathbf{B}\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{B}^T$. The problem of choosing more appropriate nonlinearities, especially $\psi$, is beyond the scope of this paper.

For all algorithms, the maximum number of iterations is 50000; the stopping criterion is $\texttt{tol} = 10^{-4}$ for OGICE and OGICENGB, $10^{-3}$ for NG and $10^{-6}$ for FICA. The step length $\mu$ was set to 0.1; 0.02 in NG; these values were selected to ensure good performance of the methods.

Figure 1 shows the success rate and median $\mathrm{SIR}_{\mathrm{imp}}$ achieved in 100 trials when the number of samples is, respectively, critical ($N = 100$) and moderate ($N = 1000$). A performance bound is given by MPDR oracle, which yields 100% success rate and 10 dB (resp. 22 dB) of median $\mathrm{SIR}_{\mathrm{imp}}$ for every $\epsilon^2$.

For $N = 100$ (row 1 in Fig. 1), NG and FICA fail to improve the initial median SIR given by MPDR ini. By contrast, OGICE and OGICENGB show higher success rate and median $\mathrm{SIR}_{\mathrm{imp}}$ than MPDR ini when $\epsilon^2 > 0.001$. OGICENGB yields significant improvements compared to OGICE, which points to its ability to exploit the non-Gaussianity of background.

The median $\mathrm{SIR}_{\mathrm{imp}}$ for $N = 1000$ (row 2, column 2 in Fig. 1) shows that the accuracy of NG and FICA is superior provided that they are initialized in a very close vicinity of the SOI ($\epsilon^2 \leq 0.01$). Here, OGICE achieves similar $\mathrm{SIR}_{\mathrm{imp}}$ to that of FICA, OGICENGB gives slightly higher $\mathrm{SIR}_{\mathrm{imp}}$ than OGICE and FICA, and NG outperforms the other methods. This is in a good agreement with the theory as NG exploits the nonGaussianity of background through separating all sources, while OGICENGB performs only a partial separation. For $\epsilon^2 > 0.1$, the median $\mathrm{SIR}_{\mathrm{imp}}$ of NG and FICA drops below $-20$ dB, which means that these algorithms mostly converge to a different source (in more than 50% of trials).

The ICE methods show superior global convergence (success rate), which is almost independent of $\epsilon^2$. Other simulations not shown here due to lack of space confirm that the global convergence of these algorithms is related to the fact that the SOI is significantly dominant in the mixture. The practical use of this interesting property will be subject of further investigations.

## 5 Conclusions and Future Works

We have shown that OGICENGB can achieve higher separation accuracy than OGICE and One-unit FastICA that assume Gaussian background. The algorithm shows excellent global convergence similarly to OGICE when the SOI is dominant, also in the scenario with a small number of samples ($N = 100$).

Open issues are the choice of a more suitable nonlinearity $\psi(\cdot)$, which might improve the accuracy of OGICENGB, and a faster optimization strategy like that of FastICA, which could considerably increase the convergence speed. Finally, the idea of this paper can be extended to the extraction of a vector component from a set of dependent instantaneous mixtures as an analogy to Independent Vector Analysis; see [9,10].

## References


1. Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. In: Proceedings of Neural Information Processing Systems, pp. 757–763 (1996)
2. Cardoso, J.F.: On the performance of orthogonal source separation algorithms. In: Proceedings of European Signal Processing Conference, pp. 776–779 (1994)
3. Comon, P.: Independent component analysis, a new concept? Sign. Process. **36**, 287–314 (1994)
4. Huber, P.J.: Projection pursuit. Ann. Statist. **13**(2), 435–475 (1985)
5. Hyvärinen, A.: One-unit contrast functions for independent component analysis: a statistical analysis. In: Proceedings of the 1997 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing VII, pp. 388–397 (1997)
6. Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Comput. **9**(7), 1483–1492 (1997)
7. Javidi, S., Mandic, D.P., Cichocki, A.: Complex blind source extraction from noisy mixtures using second-order statistics. IEEE Trans. Circuits Syst. I: Regul. Pap. **57**(7), 1404–1416 (2010)
8. Kautský, V., Koldovský, Z., Tichavský, P.: Cramér-Rao-induced bound for interference-to-signal ratio achievable through non-gaussian independent component extraction. In: 2017 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 94–97 (2017)
9. Kim, T., Attias, H.T., Lee, S.Y., Lee, T.W.: Blind source separation exploiting higher-order frequency dependencies. IEEE Trans. Audio Speech Lang. Process. **15**(1), 70–79 (2007)
10. Koldovský, Z., Tichavský, P.: Gradient algorithms for complex non-gaussian independent component/vector extraction (2018). ArXiv1803.10108
11. Koldovský, Z., Tichavský, P., Kautský, V.: Orthogonally constrained independent component extraction: Blind MPDR beamforming. In: Proceedings of European Signal Processing Conference, pp. 1195–1199 (2017)
12. Pham DTA: Blind partial separation of instantaneous mixtures of sources. In: Proceedings of International Conference on Independent Component Analysis and Signal Separation, pp. 868–875. Springer, Heidelberg (2006)
13. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech Audio Process. **12**(5), 530–538 (2004)
14. Tichavský, P., Koldovský, Z., Oja, E.: Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. IEEE Trans. Sign. Process. **54**(4), 1189–1203 (2006)
15. Tichavský, P., Koldovský, Z., Oja, E.: Corrections to "Performance analysis of the FastICA algorithm and Cramér Rao bounds for linear independent component analysis". IEEE Trans. Sign. Process. **56**(4), 1715–1716 (2008)
16. Van Trees, H.L.: Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory. Wiley, New York (2002)