Revisiting Sparse ICA from a Synthesis Point of View: Blind Source Separation for Over and Underdetermined Mixture

Fangchen Feng, Matthieu Kowalski

Revisiting Sparse ICA from a Synthesis Point of View: Blind Source Separation for Over and Underdetermined Mixtures

Fangchen Feng, Matthieu Kowalski*

Laboratoire des Signaux et Systèmes, UMR 8506 Univ Paris-Sud – CNRS – centralesupelec, 91192 Gif-sur-Yvette Cedex, France (e-mail: fangchen.feng@u-psud.fr, matthieu.kowalski@u-psud.fr).

Abstract

This paper studies the existing links between two approaches of Independent Component Analysis (ICA) – FastICA/projection pursuit and Infomax/maximum likelihood estimation – and the Sparse Component Analysis (SCA), to tackle Blind Source Separation (BSS) of the instantaneous mixtures problem. While ICA methods suit particularly well for (over)determined and noiseless mixtures, SCA has demonstrated its robustness to noise and its ability to deal with underdetermined mixtures. Using the "synthesis" point of view to reformulate ICA methods as an optimization problem, we propose a new optimization framework, which encompasses both approaches. We show that the algorithms developed to minimize the proposed functional built on SCA, but imposing a numerical decorrelation constraint on the sources, aims to improve the Signal to Inference Ratio (SIR) of the estimated sources without degrading the Signal to Distortion Ratio (SDR).

1. Introduction

The blind source separation (BSS) of instantaneous mixtures appears in various applications such as speech processing [1], biomedical processing [2] and digital communications [3]. The BSS problem is also related to the independent component analysis (ICA). This family of methods introduced in 1984 [4] has been developed to tackle the following linear problem [5]: given M observations of size T, $\mathbf{X} \in \mathbb{R}^{M\times T}$, estimate the mixing matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ and the N sources, $\mathbf{S} \in \mathbb{R}^{N\times T}$, such that

$$\mathbf{X} = \mathbf{AS} + \mathbf{E} \ , \tag{1}$$

with $\mathbf{E} \in \mathbb{R}^{M\times T}$ some additive noise.

To tackle this problem, two families of methods are mainly used: independent component analysis (ICA) [6] and sparse component analysis (SCA) [7]. The general idea of ICA is to estimate an unmixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ by minimizing a contrast function measuring the dependencies between the sources. This unmixing matrix is then used to estimate the source signals by $\mathbf{S} = \mathbf{WX}$. ICA methods have been applied with success in

*Corresponding author

a wide range of applications, such as electroencephalography (EEG) [8], functional magnetic resonance imaging (fMRI) [9, 10], and audio source separation [11]. This family of method is mainly used for noiseless (over)determined mixtures, event if some extensions exist to deal with noisy or underdetermined mixtures. One can refer to [6] for a deep presentation of ICA for BSS. The general idea of SCA is to estimate the mixing matrix $\mathbf{A}$ and the sources $\mathbf{S}$, assuming that the sources admit a sparse representation. SCA was introduced to deal with the underdetermined mixtures where the mixing matrix is first estimated, before estimating the sources [12, 13, 14]. It is also the starting point of time-frequency techniques for BSS applied to audio signals [15, 16]. SCA was also employed for (over)determined mixtures in imaging thanks to the Generalized Morphological Component Analysis (GMCA) [17], where an alternating minimization strategy is employed to estimate $\mathbf{A}$ and $\mathbf{S}$. In [18], a Bayesian model is proposed to estimate the sources in the context of SCA after providing an estimate of the (possibily underdetermined) mixing matrix, allowing sparser estimates than the classical ℓ_1 penalization used in [12, 17].

Recently, it was claimed in [19] that two of the most used ICA methods for fMRI (Infomax and FastICA, see e.g. [6]) separate sparse sources rather than independent sources, leading to the conclusion that the mathematical design of better analysis tools for brain fMRI should emphasize on other characteristics, such as sparsity, rather than independence. One given explanation is that the sparsity-based ℓ_1 minimization can be connected with InfoMax and FastICA because both of these ICA methods implicitly assume that the independent components have a generalized Gaussian distribution, which includes the sparse sources modeled by ℓ_1 minimization. This conclusion is balanced in [20] where the authors show that these two algorithms are indeed relevant to the recovery of independent fMRI sources.

Contributions and outline. As an extension of our previous work [21], we consider in this paper the BSS problem (1) in general: (over)determined and underdetermined cases, possibly with an additive white Gaussian noise. One of the main contribution of this article is to show that Infomax and FastICA can indeed be reformulated as a SCA problem. The contributions of this article are fourfold:

1. We study existing algorithms based on SCA, using a *maximum a posteriori* (MAP) approach. We propose a convergent algorithm based on proximal alternating linearized method (PALM) [22], which is more robust in practice, in the sense that it obtains acceptable results in the underdetermined case.
2. We provide a discussion on the formal links between two ICA approaches (Infomax and FastICA) and SCA approaches.
3. We propose a new framework exploiting sparsity and time decorrelation of the sources which generalizes ICA and SCA for BSS. We show that previously studied approaches are particular cases of this general formulation. Two algorithms are proposed to exploit sparsity and time decorrelation.
4. We compare all the three proposed algorithms on synthetic instantaneous audio mixtures. Even if instantaneous mixtures is not the most suitable model for audio signals [23], this framework allows to evaluate the separation results with objective measures, and subjectively by listening. The experiments show that the proposed framework outperforms existing algorithms in the underdetermined case, and can be more robust to noise in the (over)determined case. Moreover, the proposed framework is robust to the chosen number of sources to estimate.

The rest of the paper is as follow. Section 2 presents briefly the concept of ICA and SCA, as well as their state-of-the-art algorithms. In Section 3 we provide a discussion on the links between ICA and SCA, and we propose a new framework for generalizing ICA and SCA with SICA (Sparse Independant Component Analysis). We provide three new algorithms for SCA and SICA in Section 4. Experiments are done in Section 5 and Section 6 concludes the paper.

2. Blind Source Separation: ICA vs SCA

In this section, we give a brief introduction of the state-of-the-art methods based on SCA and ICA. From now and for the rest of the paper, we denote by $\{\boldsymbol{\varphi}_k \in \mathbb{R}^T\}_{k=1}^K$ a dictionary of K waveforms (such as wavelets or time-frequency atoms)[1] and $\mathbf{\Phi} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_K] \in \mathbb{R}^{T \times K}$ the associated matrix. By a slight abuse of notation, the dictionary $\{\boldsymbol{\varphi}_k \in \mathbb{R}^T\}_{k=1}^K$ will be denoted by its associated matrix $\mathbf{\Phi}$.

2.1. Sparse Component Analysis and BSS

Sparsity by analysis. Applying $\mathbf{\Phi}$ on the two sides of the general mixing model (1), it becomes

$$\mathbf{X\Phi} = \mathbf{AS\Phi} + \mathbf{E\Phi}, \tag{2}$$

$$\tilde{\mathbf{X}} = \mathbf{A}\tilde{\mathbf{S}} + \tilde{\mathbf{E}}, \tag{3}$$

where $\tilde{\mathbf{X}} = \mathbf{X\Phi}$, $\tilde{\mathbf{S}} = \mathbf{S\Phi}$ and $\tilde{\mathbf{E}} = \mathbf{E\Phi}$ are the *analysis coefficients* of $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{E}$, respectively, in the transform domain.

As mentioned in the introduction, this model is exploited in particular to deal with the underdetermined mixtures where the mixing matrix is first estimated by exploiting the sparsity of $\tilde{\mathbf{S}}$, before estimating the sources [12, 15, 13]. In [15], the authors exploit the sparsity of the sources to formulate the hypothesis of the disjointness of the sources in a time-frequency dictionary. This hypothesis has been further relaxed in [14, 32, 16]. Nevertheless, when the family of waveforms is overcomplete, several problems appear: first, if the noise $\mathbf{E}$ is assumed to be white Gaussian in the time domain, $\tilde{\mathbf{E}}$ becomes correlated in the transform domain (its density is even a degenerated normal law). Moreover, the estimated coefficients $\tilde{\mathbf{S}}$ used to synthesize the sources do not necessarily belong to the image of the operator $\mathbf{\Phi}$, thus should not be considered as analysis coefficients. That is, solving (3) without the constraint $\tilde{\mathbf{S}} = \mathbf{S\Phi}$ is not equivalent to solving (2).

Sparsity by synthesis. A simple way to deal with the drawbacks of the analysis operator is to use the synthesis modeling of the SCA [7]:

$$\mathbf{X} = \mathbf{AS} + \mathbf{E} = \mathbf{A}\boldsymbol{\alpha}\mathbf{\Phi}^* + \mathbf{E}, \tag{4}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{N \times K}$ are the *synthesis coefficients* of $\mathbf{S}$ in the dictionary $\mathbf{\Phi}$, assumed to be sparse. Using the synthesis point of view was first introduced in the seminal paper of the Basis Pursuit [33], in order to exploit sparsity. The problem is then to estimate the mixing

[1] We stick to the real case for the sake of simplicity, but the dictionary can be complex.

matrix $\mathbf{A}$ and the synthesis coefficients $\boldsymbol{\alpha}$, the sources being synthesized by $\mathbf{S} = \boldsymbol{\alpha}\boldsymbol{\Phi}^*$. Using (4), [26] proposes to estimate $\mathbf{A}$ and $\boldsymbol{\alpha}$ jointly using an alternating optimization strategy based on the MAP. Denoting by $p(\mathbf{x})$ the probability density function (pdf) of a random variable $\mathbf{x}$, the proposed MAP for BSS reads:

$$\begin{aligned} \underset{\mathbf{A},\boldsymbol{\alpha}}{\operatorname{argmax}}\ p(\mathbf{A},\boldsymbol{\alpha}|\mathbf{X}) &= \underset{\mathbf{A},\boldsymbol{\alpha}}{\operatorname{argmin}} - \log(p(\mathbf{A},\boldsymbol{\alpha}|\mathbf{X})) && (5) \\ &= \underset{\mathbf{A},\boldsymbol{\alpha}}{\operatorname{argmin}} - \log(p(\mathbf{X}|\mathbf{A},\boldsymbol{\alpha})) - \log(p(\mathbf{A})) - \log(p(\boldsymbol{\alpha})), && (6) \end{aligned}$$

where $p(\mathbf{A})$ and $p(\boldsymbol{\alpha})$ are respectively the priors of $\mathbf{A}$ and $\boldsymbol{\alpha}$.

In [26], the chosen priors are the following:

- The noise is white Gaussian: $-\log(p(\mathbf{A},\boldsymbol{\alpha}|\mathbf{X})) \propto \frac{1}{2}\|\mathbf{X} - \mathbf{AS}\|_F^2$.
- A Gaussian prior on $\mathbf{A}$: $-\log(p(\mathbf{A})) \propto \frac{\mu}{2}\|\mathbf{A}\|_F^2$, $\mu > 0$.
- A generalized Gaussian prior on the synthesis coefficients:

$$-\log(p(\boldsymbol{\alpha})) \propto \frac{\lambda}{\gamma}\sum_{n,k}|\alpha_n(k)|^\gamma \ , \quad 0 < \gamma \leq 1 \ , \ \lambda > 0 \ .$$

where $\boldsymbol{\alpha}_n(k)$ denotes the k-th element of the vector $\boldsymbol{\alpha}_n \in \mathbb{R}^K$ of the synthesis coefficients of the n-th source. $\|\cdot\|_F$ denotes the Frobenius norm. Generalized Gaussian prior with $0 < \gamma \leq 1$ allows one to favor sparse coefficients [26] and this MAP approach using sparsity is also the starting point of the generalized morphological component analysis (GMCA) [17] developed for images using wavelet basis. One of the main advantages of such an approach, is its ability to deal with the additive noise compared to simple ICA approaches (see the discussion in [17]). More complex prior can be used on the synthesis coefficients to get sparser estimate (see [18]).

2.1.1. The optimization framework

The joint estimation of the mixing matrix $\mathbf{A}$ and the synthesis coefficients $\boldsymbol{\alpha}$ in (4) can be done by solving the optimization problem of the following functional, which is equivalent to the MAP of (6) [26]:

$$\min_{\mathbf{A},\boldsymbol{\alpha}} \Gamma(\boldsymbol{\alpha},\mathbf{A}) = \min_{\mathbf{A},\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{X} - \mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2 \ + \lambda h(\boldsymbol{\alpha}) \ + g(\mathbf{A}), \tag{7}$$

where

- $\frac{1}{2}\|\mathbf{X} - \mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2$ is the data fitting term corresponding to the white Gaussian noise prior. This data fitting term will be denoted sometimes by $Q(\boldsymbol{\alpha},\mathbf{A})$ to simplify notations.
- h is the regularization term employed to favor sparse solution. A popular choice is the ℓ_1 norm [34]: $h(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1 = \sum_{n,k}|\alpha_n(k)|$.

- $\lambda > 0$ is a hyperparameter balancing between the data term and the regularization term.
- g contains constraints on $\mathbf{A}$ to avoid trivial solutions and limits the scaling ambiguity problem. A common choice for g is the indicator function of the unit circle:

  $$g(\mathbf{A}) = \imath_{\mathcal{C}}(\mathbf{A}) = \begin{cases} 0 & \text{if } \|\mathbf{a}_n\| = 1,\ n = 1, 2, \ldots, N \\ +\infty & \text{otherwise,} \end{cases} \tag{8}$$

  where $\mathbf{a}_n$ is the n-th column of $\mathbf{A}$.

The functional (7) is not convex, and then suffers from local minima. For a fixed $\mathbf{A}$, with the choice $h(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$, the functional is convex in $\boldsymbol{\alpha}$. For a fixed $\boldsymbol{\alpha}$, and denoting by $\imath_{\mathcal{B}}(\mathbf{A})$ the indicator function of the unit ball, the choice $g(\mathbf{A}) = \imath_{\mathcal{C}}(\mathbf{A})$ instead of $\imath_{\mathcal{B}}(\mathbf{A})$ makes the functional not convex. However, algorithms remain as robust as with the choice $\imath_{\mathcal{B}}(\mathbf{A})$ in practice. We stick to $g(\mathbf{A}) = \imath_{\mathcal{C}}(\mathbf{A})$ in the rest of the paper, as it appears to be the most common choice in the literature.

2.1.2. State-of-the-art algorithms

When the ℓ_1 norm is chosen to favor sparse solutions, the optimization problem (7) is non-differentiable and non-convex in $(\mathbf{A}, \boldsymbol{\alpha})$. In [26], the authors proposed to use a smooth relaxation of the ℓ_1 norm to solve the problem. However, [35] shows that the smooth technique has several drawbacks, mainly because of the choice of the smoothing parameter which balances between the convergence rate and the approximation level. It is also pointed out in [26] that the separation results are sensitive to initialization. In [36], the GMCA was developed based on an alternating optimization strategy to solve the problem for image separation in (over)determined setting. They first perform the optimization with respect to the signal and then with respect to the mixing matrix followed by a normalization step. One limitation of GMCA is that this block-coordinate descend-like algorithm does not have any convergence proof because of the extra normalization step. It is also mentioned in [36] that GMCA does not work in underdetermined case.

2.2. ICA in a nutshell

ICA was first developed in noiseless scenarios for determined mixtures ($M = N$) [6]:

$$\mathbf{X} = \mathbf{A}\mathbf{S}. \tag{9}$$

As the sources $\mathbf{S}$ are assumed to be independent, ICA methods imply the uncorrelation constraint $\mathbb{E}(\mathbf{s}(t)\mathbf{s}(\tau)) = \delta(t - \tau)$, where $\mathbf{s}(t) \in \mathbb{R}^N$ denotes the vector of sources at the time t. In practice, as the number of samples are important, this constraint leads to the numerical decorrelation: $\mathbf{S}\mathbf{S}^T = \mathbf{I}_N$, where I_N is the identity matrix of size $N \times N$. Thus, the first step of ICA method is often a whitening step on the mixtures matrix $\mathbf{X}$. In this section, we suppose that the mixture matrix verifies $\mathbf{X}\mathbf{X}^T = \mathbf{I}_M$. Consequently, in the considered ICA methods, the unmixing matrix $\mathbf{W}$ must satisfy $\mathbf{W}\mathbf{W}^T = \mathbf{W}^T\mathbf{W} = \mathbf{I}_N$.

ICA in a transform domain. A fundamental hypothesis of ICA, is that at most one source can be Gaussian [6]. In order to fulfill this assumption, it is usual to consider the mixing model (9) in a transform domain such as time-frequency or time-scale domain.

Using the dictionary $\mathbf{\Phi}$, the mixing model (9) becomes in the transform domain:

$$\begin{aligned} \mathbf{X\Phi} &= \mathbf{AS\Phi}, \\ \tilde{\mathbf{X}} &= \mathbf{A}\tilde{\mathbf{S}}, \end{aligned} \tag{10}$$

where $\tilde{\mathbf{X}} = \mathbf{X\Phi}$ and $\tilde{\mathbf{S}} = \mathbf{S\Phi}$ are the analysis coefficients of $\mathbf{X}$ and $\mathbf{S}$, respectively. ICA in the transform domain has been proposed for image processing [24] with wavelets, audio source separation with the Short Time Fourier Transform (STFT) [11] and fMRI [25] where a dictionary learning strategy is used to choose the transform domain.

We present in this subection two ICA approaches: maximum likelihood/Infomax and FastICA/projection pursuit, which are the two ICA methods discussed in [19]. We show that these two methods can be reformulated as an optimization problem to estimate $\mathbf{A}$ and $\mathbf{S}$, involving a sparse regularization term. We then propose a new functional to encompass these ICA approaches and SCA, allowing us to deal with noisy mixtures as well as underdetermined mixtures.

2.2.1. ICA based on Infomax/Maximum likelihood estimation

These methods aim to estimate $\mathbf{W}$ by maximizing the amount of mutual information or the likelihood of the model. Both lead to [6, 26]:

$$\max_{\mathbf{W}} \mathbb{E}\left(\sum_{n=1}^{N} \log\left(p_n(\mathbf{w}_n^T\mathbf{X})\right)\right) + K\log(|\det \mathbf{W}|), \tag{11}$$

with p_n the probability density function (pdf) of the independent components, under the orthogonality constraint on $\mathbf{W}$: $\mathbf{W}\mathbf{W}^T = \mathbf{W}^T\mathbf{W} = \mathbf{I}$, $\mathbf{w}_n$ being the column vector that contains the n-th row of the matrix $\mathbf{W}$.

An important point of [26] is the equivalence of their proposed optimization approach (encompassed by (7)) and the problem (11) when $\mathbf{\Phi}$ is an orthogonal basis. In such a case, the synthesis coefficients $\boldsymbol{\alpha}$ such that $\mathbf{S} = \boldsymbol{\alpha}\mathbf{\Phi}^*$ and the analysis coefficients $\tilde{\mathbf{S}}$ are equal. Then, one has

$$\tilde{\mathbf{X}} = \mathbf{A}\tilde{\mathbf{S}} = \mathbf{A}\boldsymbol{\alpha}. \tag{12}$$

Denoting by $\mathbf{W} = \mathbf{A}^{-1}$, (12) becomes

$$\mathbf{W}\tilde{\mathbf{X}} = \tilde{\mathbf{S}} = \boldsymbol{\alpha}. \tag{13}$$

Then, by re-injecting (13) into (7), the optimization becomes

$$\min_{\tilde{\mathbf{S}},\mathbf{W}} \frac{1}{2}\|\mathbf{W}\tilde{\mathbf{X}} - \tilde{\mathbf{S}}\|_F^2 \; + \lambda h(\tilde{\mathbf{S}}) \; + g(\mathbf{W}) \; . \tag{14}$$

With the choices $g(\mathbf{W}) = -K\log(|\det \mathbf{W}|)$ and $h(\tilde{\mathbf{S}}) = h(\mathbf{W}^T\tilde{\mathbf{X}}) = \mathbb{E}\left(\sum_{n=1}^{N} -\log\left(p_n(\mathbf{w}_n^T\mathbf{X})\right)\right)$, one recovers the objective (11) in the noiseless scenario $\mathbf{W}\tilde{\mathbf{X}} = \tilde{\mathbf{S}}$.

However, as already stressed in the introduction, there is no warranty to recover "true" analysis coefficients $\tilde{\mathbf{S}}$ such that $\mathbf{S}$ verifyes $\tilde{\mathbf{S}} = \mathbf{S}\boldsymbol{\Phi}$ unless if $\boldsymbol{\Phi}$ is an orthogonal basis.

2.2.2. ICA based on projection pursuit/FastICA

FastICA/projection pursuit methods aim to identify the N components of the mixture, by estimating the weight vectors $\mathbf{W}$ which maximize a measure of non-gaussianity and assuring the numerical decorrelation of the sources $\mathbf{S}\mathbf{S}^T = \mathbf{I}_N$ [6]. That is, one can reformulate ICA by projection pursuit as:

$$\max_{\mathbf{W}} \sum_n J(\mathbf{w}_n^T \mathbf{X}) \text{ s.t } \mathbf{W}\mathbf{W}^T = \mathbf{W}^T\mathbf{W} = \mathbf{I}_N, \tag{15}$$

where J is a measure of non-gaussianity. By a simple change of variable, one can reformulate (15) as:

$$\max_{\mathbf{A},\mathbf{S}} \sum_n J(\mathbf{s}_n) \text{ s.t. } \mathbf{S}\mathbf{S}^T = \mathbf{I}_N \text{ and } \mathbf{X} = \mathbf{A}\mathbf{S}. \tag{16}$$

A possible choice for J is the kurtosis of the coefficients. For a centered variable, it is equivalent to maximize:

$$J(\mathbf{S}) = \sum_n J(\mathbf{s}_n) = \sum_{n=1}^{N} \frac{\sum_t |s_n(t)|^4}{\left(\sum_t |s_n(t)|^2\right)^2} .$$

where $s_n(t) \in \mathbb{R}$ is the n-th source at the time t. Other choices are possible, such as a smooth approximation of the ℓ_0 norm as in [10] or the neg-entropy approximations used in FastICA (see [6]).

Applying FastICA in the transform domain leads to:

$$\max_{\mathbf{A},\tilde{\mathbf{S}}} J(\tilde{\mathbf{S}}) \text{ s.t. } \tilde{\mathbf{S}}\tilde{\mathbf{S}}^T = \mathbf{I}_N \text{ and } \tilde{\mathbf{X}} = \mathbf{A}\tilde{\mathbf{S}}. \tag{17}$$

One can notice that maximizing $J(\mathbf{S}) = \sum\limits_{n=1}^{N} \frac{\sum_t |s_n(t)|^4}{\left(\sum_t |s_n(t)|^2\right)^2}$ is equivalent to minimizing $\tilde{J}(\mathbf{S}) = \sum\limits_{n=1}^{N} \frac{\left(\sum_t |s_n(t)|^2\right)^2}{\sum_t |s_n(t)|^4}$ which is exactly the $\frac{\ell_p}{\ell_q}$ sparse penalty studied in [27] for deconvolution.

2.2.3. ICA and the noise

If ICA was first developed for noiseless mixtures, some versions of ICA robust to noise have been further proposed such as [28, 29]. However, these approaches seem to be less robust to noise than the SCA based methods (see the discussion in [17]).

In this article, we stick to the simple model of the white Gaussian noise. The problem of spatially correlated noise has been widely studied and can be tackled by a whitening step (see for example [30]). However, taking the spatially and temporally correlated noise into account is more complex [31]. We let the study of non white noise in inverse problems to further works, as it is out of the scope of this article dedicated to the instantaneous BSS.

3. Revisiting Sparse ICA

Using the MAP interpretation of the optimization problem (7), the regularization term h can reflect the statistical independence of the synthesis coefficients between the sources. Indeed, using the MAP approach in the Bayesian setting (6), the independence assumption reflected in the prior on $\boldsymbol{\alpha}$ leads to a separable penalty:

$$h(\boldsymbol{\alpha}) = \sum_{n=1}^{N} h_n(\boldsymbol{\alpha}_n) \ .$$

Then, in order to deal with the noise, we propose the following generalization of *Sparse ICA* (SICA):

$$\begin{cases} \min_{\mathbf{A},\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{X} - \mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2 + \sum_n h_n(\boldsymbol{\alpha}_n) + g(\mathbf{A}) \\ \text{s.t. } \mathbf{S} = \boldsymbol{\alpha}\boldsymbol{\Phi}^* \text{ and } \mathbf{S}\mathbf{S}^T = \mathbf{D}, \end{cases} \tag{18}$$

where $\mathbf{D} \in \mathbb{R}^{N\times N}$ is any diagonal matrix. The normalization constraint on the mixing matrix $\mathbf{A}$ is sufficient to solve the scaling ambiguity. In the formulation (18), we do not assume that the mixtures are whitened anymore. The decorrelation constraint on the sources then leads to the proposed constrained $\mathbf{S}\mathbf{S}^T = \mathbf{D}$, instead of $\mathbf{S}\mathbf{S}^T = \mathbf{I}_N$. As (7), the functional (18) is not convex. Moreover, (18) becomes non convex in $\boldsymbol{\alpha}$ for a fixed $\mathbf{A}$ because of the decorrelation constraint.

The next theorem shows that through the synthesis point of view, the proposed SICA (18) encompasses the FastICA/projection pursuit and Infomax/maximum likelihood techniques able to deal with noisy and underdetermined mixtures, by working on the synthesis coefficients $\boldsymbol{\alpha}$, such that $\mathbf{S} = \boldsymbol{\alpha}\boldsymbol{\Phi}^*$.

Theorem 1. *Let the BSS problem* (9)*, with* $M = N$*. Let* $\boldsymbol{\Phi}$ *be an orthogonal basis. We denote by* $\tilde{\mathbf{X}} = \mathbf{X}\boldsymbol{\Phi}$ *and* $\tilde{\mathbf{S}} = \mathbf{S}\boldsymbol{\Phi}$ *the analysis coefficients of the mixtures* $\mathbf{X}$ *and the sources* $\mathbf{S}$*, respectively. Suppose that the mixtures are whitened (*$\mathbf{X}\mathbf{X}^T = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \mathbf{I}_M$*). Let*

- $\mathbf{W}^{\mathit{Infomax}}$ *be the solution of the Infomax problem* (11) *on the analysis coefficients* $\tilde{\mathbf{X}}$*, and* $\mathbf{S}^{\mathit{Infomax}} = \mathbf{W}^{\mathit{Infomax}}\tilde{\mathbf{X}}\boldsymbol{\Phi}^*$*,*
- $\mathbf{W}^{\mathit{FastICA}}$ *be the solution of FastICA problem* (15) *on the analysis coefficients* $\tilde{\mathbf{X}}$*, and* $\mathbf{S}^{\mathit{FastICA}} = \mathbf{W}^{\mathit{FastICA}}\tilde{\mathbf{X}}\boldsymbol{\Phi}^*$*,*
- $\boldsymbol{\alpha}^{SICA,\lambda}$ *and* $\mathbf{A}^{SICA,\lambda}$ *be the solution of the sparse ICA problem* (18) *with the particular choice* $\mathbf{D} = \mathbf{I}_M$*, and* $\mathbf{S}^{SICA,\lambda} = \boldsymbol{\alpha}^{SICA,\lambda}\boldsymbol{\Phi}^*$*.*

Then, there exist g *and* h *such that*

$$\mathbf{S}^{SICA,\lambda\to 0} = \mathbf{S}^{\mathit{Infomax}},$$

or

$$\mathbf{S}^{SICA,\lambda\to 0} = \mathbf{S}^{\mathit{FastICA}}.$$

Proof. First, the orthogonality constraint on $\mathbf{W}$: $\mathbf{W}\mathbf{W}^T = \mathbf{I}_M$ as well as the hypothesis $\mathbf{X}\mathbf{X}^T = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \mathbf{I}_M$ in ICA implies that $\mathbf{S}\mathbf{S}^T = \mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T = \mathbf{I}_M$. The second ingredient of the proof is that $\mathbf{\Phi}$ being an orthogonal basis, we have $\boldsymbol{\alpha}_n = \tilde{\mathbf{s}}_n = \mathbf{w}_n^T\tilde{\mathbf{X}}$.

Let $g(\mathbf{A}) = K\log(|\det\mathbf{A}|) = -K\log(|\det\mathbf{W}|)$ and $h = \sum_{n=1}^{N} -\log p_n$, with p_n defined by Eq. (11). Then, when $\lambda \to 0$, SICA (18) becomes:

$$\begin{aligned} &\min_{\mathbf{A},\boldsymbol{\alpha}} K\log(|\det\mathbf{A}|) - \sum_{n=1}^{N}\log p_n(\boldsymbol{\alpha}_n) \\ &\text{s.t. } \mathbf{X} = \mathbf{A}\boldsymbol{\alpha}\mathbf{\Phi}^* \text{ and } \mathbf{S}\mathbf{S}^T = \mathbf{I}_M. \end{aligned}$$

SICA can then be rewritten as the optimization problem on $\mathbf{W}$:

$$\begin{aligned} &\max_{\mathbf{W}} K\log(|\det\mathbf{W}|) + \sum_{n=1}^{N}\log p_n(\mathbf{w}_n^T\tilde{\mathbf{X}}) \\ &\text{s.t. } \mathbf{W}\mathbf{W}^T = \mathbf{I}_M, \end{aligned}$$

which proves the first point of the theorem.

Let $g(\mathbf{A}) = \imath_{\mathcal{C}}(\mathbf{A})$ and $h(\boldsymbol{\alpha}) = -J(\boldsymbol{\alpha})$ with J defined by Eq. (17). Then, when $\lambda \to 0$, SICA (18) becomes:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \; -J(\boldsymbol{\alpha}) \\ &\text{s.t. } \mathbf{X} = \mathbf{A}\boldsymbol{\alpha}\mathbf{\Phi}^* \text{ and } \mathbf{S}\mathbf{S}^T = \mathbf{I}_M. \end{aligned}$$

SICA can then be rewritten as the optimization problem on $\mathbf{W}$:

$$\begin{aligned} &\max_{\mathbf{W}} J(\mathbf{W}\tilde{\mathbf{X}}) \\ &\text{s.t. } \mathbf{W}\mathbf{W}^T = \mathbf{I}_M, \end{aligned}$$

which concludes the proof.

When $\mathbf{\Phi}$ is overcomplete, we can only conclude that, for $\lambda \to 0$

$$\Gamma(\boldsymbol{\alpha}^{\text{SCA}}, \mathbf{A}^{\text{SCA}}) \leq \Gamma(\tilde{\mathbf{S}}^{\text{FastICA}}, \mathbf{W}^{\text{FastICA}^{-1}}),$$

and

$$\Gamma(\boldsymbol{\alpha}^{\text{SCA}}, \mathbf{A}^{\text{SCA}}) \leq \Gamma(\tilde{\mathbf{S}}^{\text{InfoMax}}, \mathbf{W}^{\text{InfoMax}^{-1}}).$$

The connections between ICA and SCA appear several times. In [24], the authors have remarked that the contrast function employed in ICA can be interpreted as a measure of sparsity. In [37], it is shown that ICA methods work better in a *transform domain* such as Curvelets or Ridgelets, and the authors justify the use of kurtosis in ICA by a sparse coding point of view. This remark was already made in [24] at the end of 90's. In [12, 13], the estimation of the mixing matrix for underdetermined mixtures is performed by exploiting the sparsity of the sources in the transform domain, and the independence can be viewed as a consequence of sparsity. In [17], the authors already discussed some existing links between ICA and SCA. In particular, they show that the ℓ_1 regularizer is indeed a contrast function for ICA. We have here generalized and summarized all these results to clearly show that FastICA and Infomax *are* actually sparse component analysis methods with a decorrelation constraint to solve the blind source separation problem, when $M = N$.

In the rest of this paper, for the sake of simplicity, we stick to ℓ_1 norm for the synthesis coefficients. Therefore we consider the following formulation:

$$\begin{cases} \min_{\mathbf{A},\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{X} - \mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_1 + \imath_{\mathcal{C}}(\mathbf{A}) \\ \text{s.t. } \boldsymbol{\alpha}\boldsymbol{\Phi}^*\boldsymbol{\Phi}\boldsymbol{\alpha}^* = \mathbf{D}. \end{cases} \tag{19}$$

where $\mathbf{D} \in \mathbb{R}^{N\times N}$ is any diagonal matrix.

4. Algorithms for SICA

In this section, we first present a simple convergent algorithm for SCA, by applying the proximal alternating linearized method (PALM) [22]. Then, we propose two algorithms to solve the problem (18). We use an alternating direction method of multipliers (ADMM) approach before providing a simplified version, which appears to be faster and more robust in practice.

4.1. A convergent algorithm for SCA: BSS-PALM

Applied to (7) with $h(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$, the PALM leads to a simple algorithm using a gradient descent step on the data term followed by a thresholding step to estimate the sources coefficients $\boldsymbol{\alpha}$, $\mathbf{A}$ being fixed, and a gradient descent step on a data term followed by a projection on the unit circle to estimate the mixing matrix $\mathbf{A}$, $\boldsymbol{\alpha}$ being fixed. The algorithm is described in Alg. 1, where we use

- The soft-thresholding operator:

$$\mathcal{S}_\lambda(\mathbf{x}) = \mathbf{x}\left(1 - \frac{\lambda}{|\mathbf{x}|}\right)^+ , \tag{20}$$

  where the multiplication and division are applied element-wise, as well as the operator $(x)^+ = \max(0, x)$.

- The normalization of each column of the matrix $\mathbf{A}$

$$\tilde{\mathbf{A}} = \mathcal{P}_{\mathcal{C}}(\mathbf{A}), \tag{21}$$

  such that each column $\tilde{\mathbf{a}}_n$ of $\tilde{\mathbf{A}}$ is normalized: $\tilde{\mathbf{a}}_n = \frac{\mathbf{a}_n}{\|\mathbf{a}_n\|}$, $\forall n \in \{1, \ldots, N\}$.

The derivation of this algorithm is detailed in Appendix A.1.

Using a direct application of PALM [22, Theorem 3.1] to solve (7), one obtains directly the next proposition.

Algorithm 1: BSS-PALM

```
Initialization : $\boldsymbol{\alpha}^{(1)} \in \mathbb{R}^{N\times K}$, $\mathbf{A}^{(1)} \in \mathbb{R}^{M\times N}$, $L^{1,(1)} = \|\mathbf{A}^{(1)}\|_2^2$, $L^{2,(1)} = \|\boldsymbol{\alpha}^{(1)}\boldsymbol{\Phi}^*\|_2^2$,
$j = 1$;
repeat
    1. $\nabla_{\boldsymbol{\alpha}} Q\left(\boldsymbol{\alpha}^{(j)}, \mathbf{A}^{(j)}\right) = -\mathbf{A}^{(j)^T}\left(\mathbf{X} - \mathbf{A}^{(j)}\boldsymbol{\alpha}^{(j)}\boldsymbol{\Phi}^*\right)\boldsymbol{\Phi}$;
    2. $\boldsymbol{\alpha}^{(j+1)} = \mathcal{S}_{\lambda/L^{1,(j)}}\left(\boldsymbol{\alpha}^{(j)} - \frac{1}{L^{1,(j)}}\nabla_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}^{(j)}, \mathbf{A}^{(j)})\right)$;
    3. $\nabla_{\mathbf{A}} Q(\boldsymbol{\alpha}^{(j+1)}, \mathbf{A}^{(j)}) = -(\mathbf{X} - \mathbf{A}^{(j)}\boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*)\boldsymbol{\Phi}\boldsymbol{\alpha}^{(j+1)^H}$;
    4. $\mathbf{A}^{(j+1)} = \mathcal{P}_{\mathcal{C}}\left(\mathbf{A}^{(j)} - \frac{1}{L^{2,(j)}}\nabla_{\mathbf{A}} Q(\boldsymbol{\alpha}^{(j+1)}, \mathbf{A}^{(j)})\right)$;
    5. $L^{1,(j+1)} = \|\mathbf{A}^{(j+1)}\|_2^2$;
    6. $L^{2,(j+1)} = \|\boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*\|_2^2$;
    7. $j = j + 1$;
until convergence;
```

Proposition 1. *The sequence* $(\mathbf{A}^{(j)}, \boldsymbol{\alpha}^{(j)})$ *generated by Alg. 1 converges to a critical point of problem:*

$$\min_{\mathbf{A},\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{X} - \mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2 \ + \lambda\|\boldsymbol{\alpha}\|_1 + \imath_{\mathcal{C}}(\mathbf{A}).$$

4.2. ADMM approach

We first reformulate problem (19) with a linear constraint by introducing an extra variable as follows:

$$\begin{cases} \underset{\mathbf{A},\boldsymbol{\alpha},\mathbf{S}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{X} - \mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*||_F^2 + \lambda||\boldsymbol{\alpha}||_1 + \imath_{\mathcal{C}}(\mathbf{A}) + \imath_{\text{deco}}(\mathbf{S}) \\ \text{s.t. } \mathbf{S} = \boldsymbol{\alpha}\boldsymbol{\Phi}^*, \end{cases} \tag{22}$$

where $\imath_{\text{deco}}$ is an indicator function of the decorrelation constraint of $\mathbf{S}$, reading:

$$\imath_{\text{deco}}(\mathbf{S}) = \begin{cases} 0 & \text{if } \mathbf{S}\mathbf{S}^T = \mathbf{D}\text{, with } \mathbf{D} \text{ being diagonal} \\ +\infty & \text{otherwise.} \end{cases} \tag{23}$$

We then apply the linearized preconditioned alternating direction method of multipliers (LPADMM) [38] to solve the problem (22).

The general idea of ADMM is based on the alternating optimization of the corresponding augmented Lagrangian function:

$$\begin{aligned} L(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{S}, \boldsymbol{\eta}) =& \frac{1}{2}||\mathbf{X} - \mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*||_F^2 + \lambda||\boldsymbol{\alpha}||_1 + \imath_{\mathcal{C}}(\mathbf{A}) + \imath_{\text{deco}}(\mathbf{S}) + \langle \boldsymbol{\eta}, \mathbf{S} - \boldsymbol{\alpha}\boldsymbol{\Phi}^*\rangle + \frac{\gamma}{2}\|\mathbf{S} - \boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2 \\ =& F(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{S}, \boldsymbol{\eta}) + \imath_{\mathcal{C}}(\mathbf{A}) + \imath_{\text{deco}}(\mathbf{S}), \end{aligned}$$

with

$$F(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{S}, \boldsymbol{\eta}) = \frac{1}{2}||\mathbf{X} - \mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*||_F^2 + \langle \boldsymbol{\eta}, \mathbf{S} - \boldsymbol{\alpha}\boldsymbol{\Phi}^*\rangle + \frac{\gamma}{2}\|\mathbf{S} - \boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2 \ .$$

The linearized and preconditioned version (LPADMM) is intended to simplify the algorithm, that is, in each iteration, instead of minimizing the original function, to minimize its first order approximation. ADMM approach has been widely used for inverse problem using sparsity (see for example [39]). While used in practice for some non-convex problems (for example [40]), the convergence of ADMM algorithms in a non-convex setting is currently under study [41, 42]. The derivation of this algorithm is postponed in Appendix A.2. We refer to it as BSS-LPADMM in the following and summarise it in Alg. 2.

Algorithm 2: BSS-LPADMM

```
Initialization : $\boldsymbol{\alpha}^{(1)} \in \mathbb{R}^{N\times K}$, $\mathbf{S}^{(1)} \in \mathbb{R}^{N\times T}$, $\mathbf{A}^{(1)} \in \mathbb{R}^{M\times N}$, $L^{F,(1)} = \|\mathbf{A}^{(1)}\|_2^2 + \gamma$,
$L^{(1)} = \|\boldsymbol{\alpha}^{(1)}\boldsymbol{\Phi}^*\|_2^2$, $j = 1$;
repeat
    1. $\nabla_{\boldsymbol{\alpha}} F(\mathbf{A}^{(j)}, \boldsymbol{\alpha}^{(j)}, \mathbf{S}^{(j)}, \boldsymbol{\eta}^{(j)}) =$
       $-\mathbf{A}^{(j)^T}(\mathbf{X} - \mathbf{A}^{(j)}\boldsymbol{\alpha}^{(j)}\boldsymbol{\Phi}^*)\boldsymbol{\Phi} - \boldsymbol{\eta}^{(j)}\boldsymbol{\Phi} - \gamma(\mathbf{S}^{(j)} - \boldsymbol{\alpha}^{(j)}\boldsymbol{\Phi}^*)\boldsymbol{\Phi}$;
    2. $\boldsymbol{\alpha}^{(j+1)} = \mathcal{S}_{\lambda/L^{F,(j)}}(\boldsymbol{\alpha}^{(j)} - \frac{1}{L^{F,(j)}}\nabla_{\boldsymbol{\alpha}} F(\mathbf{A}^{(j)}, \boldsymbol{\alpha}^{(j)}, \mathbf{S}^{(j)}, \boldsymbol{\eta}^{(j)}))$;
    3. $\mathbf{S}^{(j+1/2)} = \boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^* - \boldsymbol{\eta}^{(j)}/\gamma$;
    4. $\Sigma_{\mathbf{S}^{(j+1/2)}} = \mathbf{S}^{(j+1/2)}\mathbf{S}^{(j+1/2)^T}$;
    5. $\mathbf{W}_{\mathbf{S}^{(j+1/2)}} = (\mathrm{diag}(\Sigma_{\mathbf{S}^{(j+1/2)}}))^{1/2}\Sigma_{\mathbf{S}^{(j+1/2)}}^{-1/2}$;
    6. $\mathbf{S}^{(j+1)} = \mathbf{W}_{\mathbf{S}^{(j+1/2)}}\mathbf{S}^{(j+1/2)}$;
    7. $\nabla_{\mathbf{A}} Q(\mathbf{A}^{(j)}, \boldsymbol{\alpha}^{(j+1)}) = -(\mathbf{X} - \mathbf{A}^{(j)}\boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*)\boldsymbol{\Phi}\boldsymbol{\alpha}^{(j+1)^*}$;
    8. $\mathbf{A}^{(j+1)} = \mathcal{P}_{\mathcal{C}}\left(\mathbf{A}^{(j)} - \frac{1}{L^{(j)}}\nabla_{\mathbf{A}} Q(\mathbf{A}^{(j)}, \boldsymbol{\alpha}^{(j+1)})\right)$;
    9. $\boldsymbol{\eta}^{(j+1)} = \boldsymbol{\eta}^{(j)} + \gamma(\mathbf{S}^{(j+1)} - \boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*)$;
    10. $L^{F,(j+1)} = \|\mathbf{A}^{(j+1)}\|_F^2 + \gamma$;
    11. $L^{(j+1)} = \|\boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*\|_F^2$;
    12. $j = j + 1$;
until convergence;
```

4.3. A simplified version

The BSS-LPADMM algorithm solves directly the problem (19) but is subject to big computational burden. Therefore, we design here a simplified version in Alg. 3 and refer to this algorithm as BSS-Deco.

Compared to BSS-LPADMM, we set the dual variable $\boldsymbol{\eta}$ and the penalty parameter γ to zero. Despite the lack of convergence proof, experiments support its good performance.

4.4. Determining the number of sources

The authors of [43] show that if one has only an upper bound of the number of sources, acceptable estimation of the source signals can still be obtained by the analysis sparsity minimization with ℓ_1 norm. They showed that the extra source channels will contain little energy thus do not have obvious negative effect on the source estimation. We employed the simple strategy presented in Alg. 4 to eliminate extra source channels during the iterations, where I_M and I_m are the source index which correspond to the source of

Algorithm 3: BSS-Deco

```
Initialisation : $\boldsymbol{\alpha}^{(1)} \in \mathbb{R}^{N\times K}$, $\mathbf{A}^{(1)} \in \mathbb{R}^{M\times N}$, $L^{(1)} = \|\mathbf{A}^{(1)}\|_2^2$, $j = 1$;
repeat
    1. $\nabla_{\boldsymbol{\alpha}} Q(\mathbf{A}^{(j)}, \boldsymbol{\alpha}^{(j)}) = -(\mathbf{X} - \mathbf{A}^{(j)}\boldsymbol{\alpha}^{(j)}\boldsymbol{\Phi}^*)\boldsymbol{\Phi}\boldsymbol{\alpha}^{(j)^*}$;
    2. $\boldsymbol{\alpha}^{(j+1)} = \mathcal{S}_{\lambda/L^{(j)}}(\boldsymbol{\alpha}^{(j)} - \frac{1}{L^{(j)}}\nabla_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}^{(j)}, \mathbf{A}^{(j)}))$;
    3. $\mathbf{S}^{(j+1/2)} = \boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*$;
    4. $\Sigma_{\mathbf{S}^{(j+1/2)}} = \mathbf{S}^{(j+1/2)}\mathbf{S}^{(j+1/2)^T}$;
    5. $\mathbf{W}_{\mathbf{S}^{(j+1/2)}} = (\mathrm{diag}(\Sigma_{\mathbf{S}^{(j+1/2)}}))^{1/2}\Sigma_{\mathbf{S}^{(j+1/2)}}^{-1/2}$;
    6. $\mathbf{S}^{(j+1)} = \mathbf{W}_{\mathbf{S}^{(j+1/2)}}\mathbf{S}^{(j+1/2)}$;
    7. $\mathbf{A}^{(j+1)} = \mathcal{P}_{\mathcal{C}}(\mathbf{X}\mathbf{S}^{(j+1)^T})$;
    8. $L^{(j+1)} = \|\mathbf{A}^{(j+1)}\|_2^2$;
    9. $j = j + 1$;
until convergence;
```

maximum and minimum energy respectively. E_{I_M} and E_{I_m} are the corresponding energy, and ϵ a threshold.

Algorithm 4: BSS-Deco with the number of sources determination

```
repeat
    1. Update $\boldsymbol{\alpha}$, $\mathbf{S}$ and $\mathbf{A}$ according to one iteration of Algo. 3;
    2. if $E_{I_M}/E_{I_m} > \epsilon$ then
        Eliminate $\mathbf{s}_{I_m}$, $\boldsymbol{\alpha}_{I_m}$ and $\mathbf{a}_{I_m}$
    3. $j = j + 1$;
until convergence;
```

5. Experiments

After presenting the experimental setup, we discuss the choice of the hyperparameter λ and the robustness to the choice of the number of unknown sources. We then compare all the proposed algorithms and the state-of-the-art ICA and SCA algorithms on over/underdetermined mixtures with and without additive white Gaussian noise.

Results and matlab code are available on `http://fcfeng28.wixsite.com/monsite/professional-use`.

5.1. Experimental setup

The algorithms are evaluated on mixtures created with 10 sets of signals used in [44], issued from the SiSEC2011 database [45], with a sample rate at 11 kHz and a duration of 6 s. All the sources have the same order of energy. The mixing matrix was generated randomly following a normal distribution with normalized columns. The STFT was computed with half-overlapping tight Hann window of 512 samples length (about 46.5 ms) using the LTFAT toolbox [46]. All the proposed algorithms are initialized randomly following a normal distribution. Despite the non convexity of the approaches, we have observed that the results are robust to this initialization.

The separation performances were assessed using the Signal to Distortion Ratio (SDR) and Signal to Interference Ratio (SIR) [47]. The SDR indicates the overall quality of each estimated source compared to the target, while the SIR reveals the amount of residual crosstalk from the other sources. A larger value of SDR/SIR means a better quality of separation.

To show the improvement brought by the perfect knowledge of the mixing matrix, we design two "non blind" oracle settings for the proposed algorithms for comparison. These two oracles are denoted by BSS-Oracle (corresponding to the non blind version of BSS-PALM and BSS-Deco) and BSS-LPADMM-Oracle (corresponding to the non blind version of BSS-LPADMM-Oracle).

Finally, the parameter γ for BSS-LPADMM was set empirically to $\gamma = 0.05$.

5.2. Choice of the hyperparameter λ

Several methods have been studied for automatic choice of λ in inverse problem, such as projected GSURE (generalized Stein unbiased risk estimator, see [48] and references therein) or the Stein Unbiased GrAdient estimator of the Risk (SUGAR) [49]. However, most of the proposed methods imply to compute several solutions for several λ, and then choose the "best" solution according to some criteria. If such a blind method is needed for some applications, it can also be required to let the user decide what is the "best" solution. Particularly for signal (audio, image, video...) restoration, the best acceptable result will not always fit any "objective" criteria. Such a discussion can be found for example in [50] for audio signals.

We stick in this article to simple choices for the hyperparameter:

- If the mixture is assumed to be free of noise, we choose $\lambda \to 0$ in order to avoid any "denoising" on the estimation. In practice, for small value of λ, we used the continuation trick also known as warm-start or fixed-point continuation [51]: we first run the algorithm with a large value of λ, and then iteratively decreased the parameter till the wanted value.
- If some noise is added, then we choose the λ giving the best results in term of SDR for each algorithm. This choice is rarely the best choice from a subjective point of view, and cannot be automatically done in practice. However, it appears to be the most fair, by giving the best achievable result from a SDR point of view (the SDR giving the overall quality of each estimated source compared to the target).

In addition to this two "default" choices, we also provide a short discussion about the influence of λ on the results.

5.3. Robustness to the number of sources

The following experiments show that the proposed algorithms are robust to the number of sources, using Alg. 4 with the threshold ϵ empirically fixed to 3.5 in our experiments. For the sake of simplicity, we only show the results obtained in the noiseless setting for the proposed BSS-Deco in Alg. 3. Other algorithms lead to similar results. In these experiments, the number of microphones was set to $M = 3$, the real number of sources N_r and its upper bound N varied from 2 to 5. Fig. 1 displays the separation results and supports the robustness of the proposed algorithm to the choice of the number of sources. One can observe that the curves are almost constant, which mean that an over-estimation of the sources does not decrease the performances: the right number of sources is selected by Alg. 4.

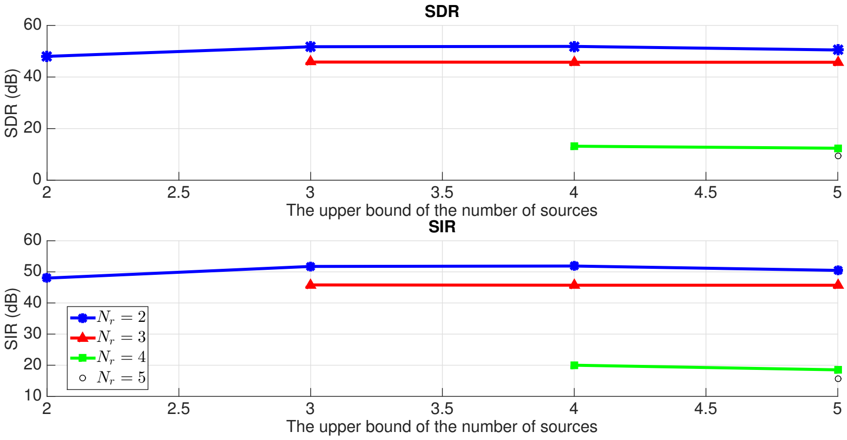

Figure 1: Performances as a function of the upper bound of number of sources for different cases

5.4. Overdetermined BSS

In this setting, the number of sources is fixed to $N = 3$ and the number of microphones runs from 3 to 10 (3, 5, 7 and 10).

5.4.1. Noiseless case

As reference for ICA approaches, we provide the results obtained by Efficient FastICA (EFICA) [52], an improved version of FastICA whose residual error variance attains asymptotically the Cramér-Rao lower bound, and second order blind identification (SOBI) [53] using the toolbox [54].

In the noiseless case, the whitening and dimension reduction used in ICA is justified. We therefore present the results obtained by all the algorithms –except for ICA approaches– with and without this pre-processing step. The results are summarized in Table 1.

It is clear that the whitening pre-processing step greatly improves the results of BSS-PALM and GMCA, which outperform other approaches, while their performances are the worst without this pre-processing step. Moreover, it is also interesting to note that the performances of all the sparsity-based algorithms improve with the number of observations when such a pre-processing step is not used.

This experiment shows that the proposed algorithms are able to well separate overdetermined mixtures as expected. In this case, the ICA approaches are indeed the most

Table 1: Performances of different algorithms in noiseless (over)determined setting for $N = 3$ (SDR / SIR). On a line, the best performance is in black bold. If the difference between a performance and the best is less than 1 dB, it is displayed in gray bold.

	BSS-PALM	BSS-ADMM	BSS-Deco	GMCA	EFICA	SOBI
$M = 3$	17.0 / 17.0	41.6 / 41.6	**45.8 / 45.8**	17.8 / 17.8	-	-
$M = 5$	19.4 / 19.4	44.3 / 44.3	**46.0 / 46.0**	18.0 / 18.0	-	-
$M = 6$	25.6 / 25.6	**45.7 / 45.7**	**46.1 / 46.1**	22.8 / 22.8	-	-
$M = 7$	29.9 / 29.9	**45.6 / 45.6**	**46.9 / 46.9**	27.5 / 27.5	-	-
$M = 10$	39.3 / 39.2	**48.2 / 48.2**	**47.6 / 47.9**	38.3 / 38.3	-	-
With whitening	**66.0 / 66.0**	48.4 / 48.4	48.3 / 48.3	**66.6 / 68.4**	49.5 / 49.5	37.3 / 37.3

appropriate choice to separate such mixtures thanks to the simplicity and rapidity of the practical algorithms (cf. Sec. 5.6 about the computational time).

5.4.2. Noisy case

Performance as a function of input Signal to Noise Ratio (SNR). In this experiment, we stick to the determined setting, i.e. $M = 3, N = 3$. The results are summarized on Fig. 2.

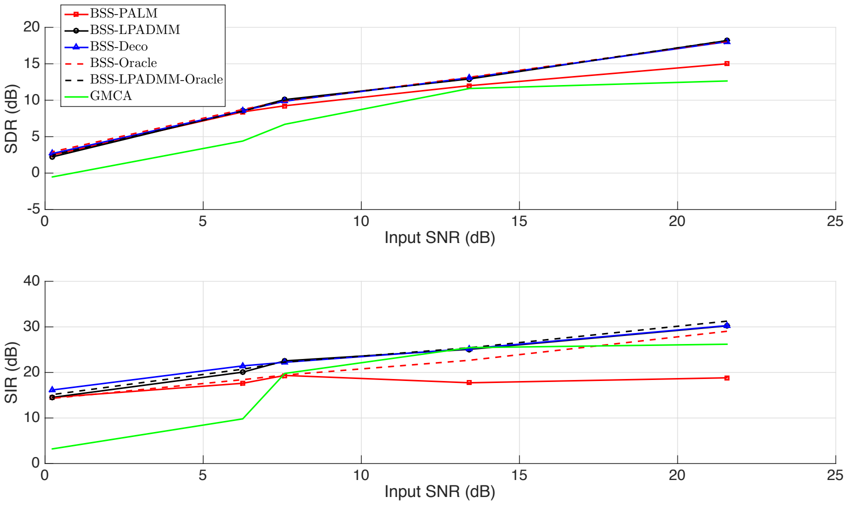

Figure 2: Performances as a function of the input SNR in determined setting ($M = 3$, $N = 3$)

Once again, BSS-LPADMM and BSS-Deco have similar performance in terms of SDR and outperform BSS-PALM especially when the input SNR is relatively high. In terms of SIR, as expected, BSS-LPADMM and BSS-Deco lead to the best performances, comparable with BSS-LPADMM-Oracle.

One of the most remarkable results, is that the two oracle settings perform similarly as their corresponding "blind" algorithms in terms of SDR, while BSS-PALM and BSS-Oracle obtain the two worst SIR. This last point supports the intuition that the decorrelation constraint is particularly important to improve the SIR. These results support the conclusion of [55], where SCA approaches appears to be the most robust to noise.

Performance as a function of the number of observations. In this experiment, a white Gaussian noise is added to reach an input SNR of 7.58 dB. The results are summarized on Fig. 3.

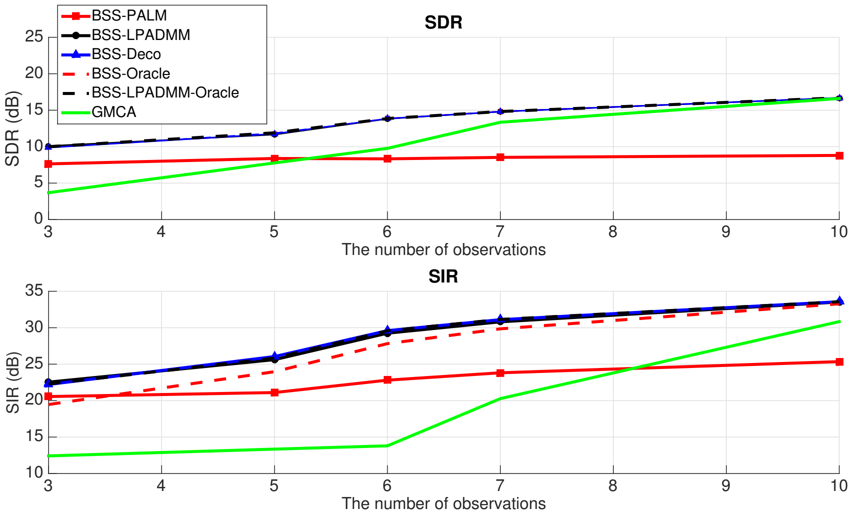

Figure 3: Performances as a function of the number of microphones in (over)determined noisy scenario with $N = 3$ and input SNR equaling 7.58 dB

Similar remarks as previous paragraph can be made: BSS-LPADMM and BSS-Deco reach the best performances, comparable to BSS-LPADMM Oracle, for both SDR and SIR, while BSS-Oracle shows bad performances in term of SIR. As expected, SDR and SIR grows with the number of observations for all algorithms.

Performance as a function of sparsity level. As mentioned before, the hyperparameter λ in the proposed algorithms is linked to the variance of the input noise and controls the sparsity level of the estimated sources. Therefore, we present the performances of the proposed algorithms as a function of the sparsity level[2] of the estimated source signals on Fig. 4.

In this case, the behavior of the SDR and SIR are comparable. We can notice that, for BSS-LPADMM and BSS-Deco, the best performance is obtained when the sparsity level is around 85%. Empirically, this sparsity level corresponds to $\lambda \simeq \sigma$ where σ is the standard deviation of the input noise.

5.5. Underdetermined BSS

In these experiments, the number of microphones varies from 2 to 5 and the number of sources runs from 3 to 6. We compare the proposed algorithms to the state-of-the art approaches where the mixing matrix is first estimated using Demix [14], then the sources are estimated by the time-frequency masking (DUET) [15] or the ℓ_1 minimization of the

[2] Sparsity level here means the percentage of zero values in the vector or matrix.

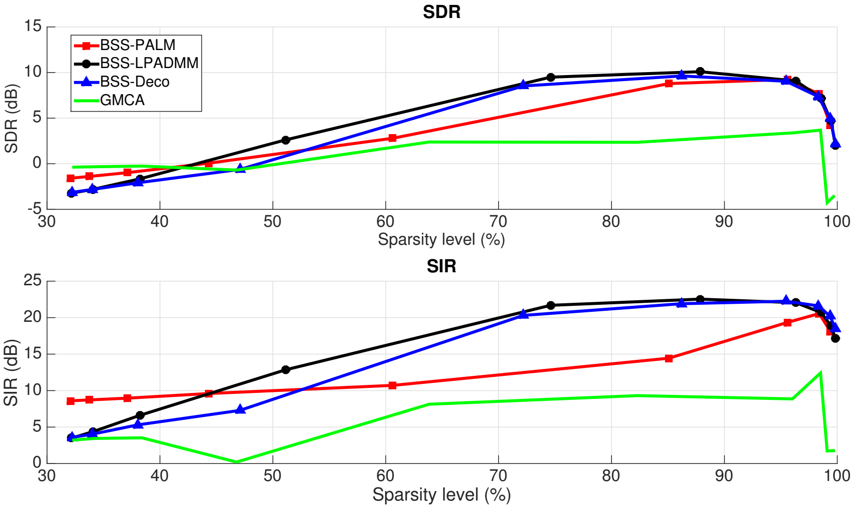

Figure 4: Performances as a function of the sparsity level of the estimated sources for SNR=7.58 dB in determined scenario ($M = 3$, $N = 3$)

analysis coefficients [16]. We denote these two methods by Demix-DUET and Demix-ℓ_1 respectively. We must stress that, in practice, Demix cannot work when the number of sources is larger than 5 with the number of observations $M = 2$.

5.5.1. Noiseless case

Performance as a function of the number of observations. We first fix the number of sources to $N = 6$. Table 2 shows the evolution of SDR and SIR with respect to the number of observations. As DUET method is mainly for two-microphone setting, its performance is not shown.

Table 2: Performances of different algorithms in noiseless underdetermined setting with the number of sources $N = 6$ (SDR / SIR). On a line, the best performance is in black bold. If the difference between a performance and the best is less than 1 dB, it is displayed in gray bold.

	BSS-PALM	BSS-ADMM	BSS-Deco	BSS-Oracle	BSS-LPADMM-Oracle	Demix-ℓ_1
$M = 2$	01.4 / 05.9	01.9 / 06.2	01.3 / 07.8	**03.0** / 07.9	**02.5** / **09.2**	-
$M = 3$	**07.5** / 11.5	06.4 / 12.6	**07.5** / **14.3**	**07.9** / 12.5	06.7 / 13.0	**08.0** / 12.1
$M = 4$	**12.9** / 17.0	11.8 / 18.1	**12.9** / **19.7**	**13.3** / 18.0	12.1 / 18.6	**13.2** / 17.1
$M = 5$	15.3 / 17.1	17.7 / 24.1	**20.1** / **26.3**	**20.2** / 24.7	18.7 / 24.6	**20.1** / 24.1

Except for $M = 2$, the Demix-ℓ_1 reaches the best SDR among all the non-oracle algorithms, but the difference between BSS-Deco is less than 1 dB, while BSS-Deco outperforms other approaches in terms of SIR. One can remark that BSS-LPADMM-Oracle is also outperformed by BSS-Deco, which can be explained by the fact that BSS-LPADMM can be sensitive to local minima.

Performance as a function of the number of sources. In these experiments, the number of observations is fixed to $M = 2$. We provide on Fig. 5 the evolution of SDR and SIR with respect to the number of sources.

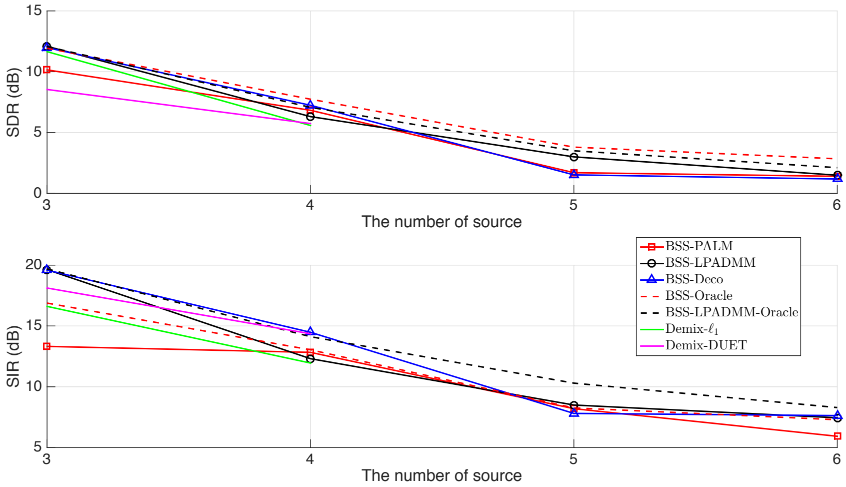

Figure 5: Performances as a function of the number of sources in underdetermined noiseless scenario ($M = 2$)

As expected, the performances collapse when N grows. For $N \leq 4$, all approaches except BSS-PALM and Demix-DUET provide comparable results in term of SDR. BSS-Deco reaches the best SIR (greater than BSS-Oracle and close to BSS-LPADMM-Oracle). For $N > 5$, performances of BSS-Deco collapse in terms of SDR, but still outperforms other non-oracle algorithms in terms of SIR.

5.5.2. Noisy case

Performance as a function of the input SNR. Fig. 6 displays the separation results of the proposed algorithms as a function of the input SNR, with the number of sources fixed to $N = 3$ and the number of observations to $M = 2$.

Once again, the two oracle algorithms outperform the others in terms of SDR, while only BSS-LPADMM-Oracle outperforms other approaches in terms of SIR, still supporting the fact that taking decorrelation into account improves the SIR of the estimated source signals.

From a SDR point view, all algorithms are comparable, except Demix-DUET. BSS-PALM and Demix-ℓ_1 perform a little worse, but the difference is less than 1 dB. The major difference between the algorithms is from a SIR point of view: BSS-LPADMM and BSS-Deco clearly outperform other approaches, including BSS-Oracle. Demix algorithm does not work when the input SNR is less than 9 dB.

Performance as a function of the sparsity level. As in the (over)determined case, we present in Fig. 7 the separation performances as a function of the sparsity level of the estimated source signals for SNR=23.43 dB and 9 dB.

We can see on Fig. 7 that a compromise must be performed between the SDR and the SIR in the underdetermined case: a small improvement on the SDR can lead to a

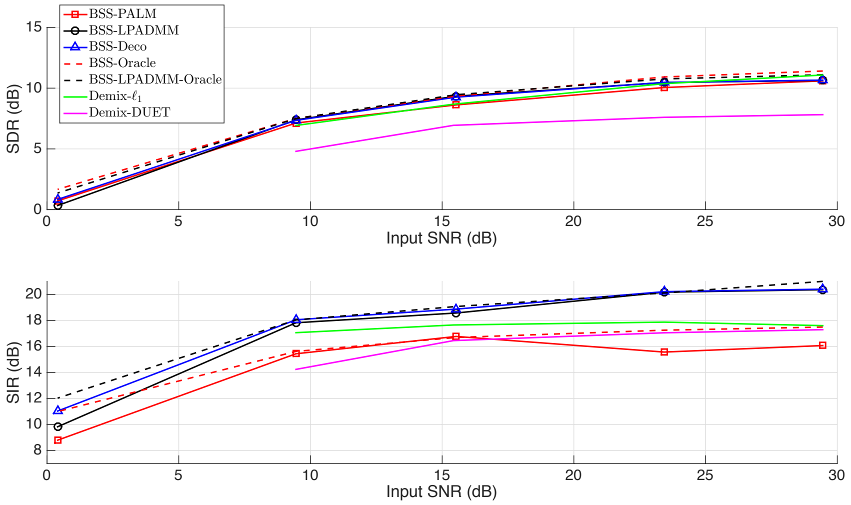

Figure 6: Performances as a function of the input SNR in underdetermined scenario ($M = 2$, $N = 3$)

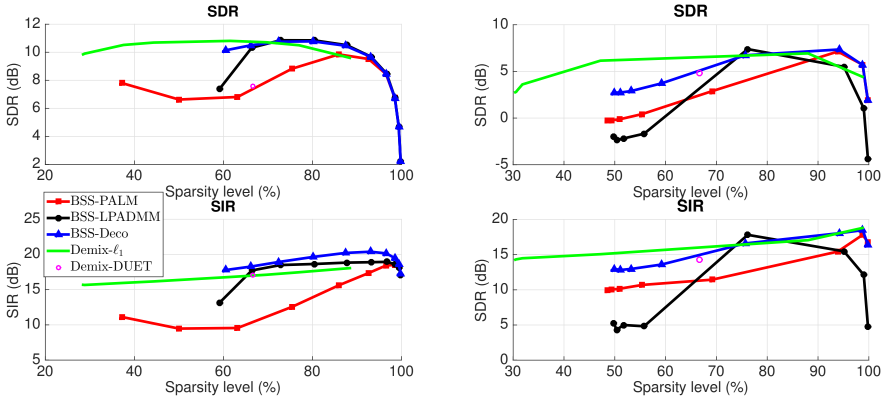

Figure 7: Performances as a function of the sparsity level in underdetermined noisy case ($M = 2$, $N = 3$) with SNR=23.4 dB (left) and SNR=9 dB (right)

big decrease of the SIR. This behavior differs from the similar experiments performed in the determined case in Fig 4.

5.6. Computational comparison

We end the experiment section by giving some indications about the computational time of different algorithms. Table 3 shows the computational time for the previously mentioned sparsity-based algorithms with 20000 iterations, which is the number of iterations used in practice for the experiments. The computational time of EFICA and SOBI using ICALAB toolbox and Demix-DUET method is about 1 second. In Fig. 8, we show the evolution of the SDR with respect to the CPU time, in the underdetermined noiseless scenario of Sec. 5.5.1. The computational time of GMCA with $M = 3$ and $N = 3$ is 1600 s for 20000 iterations.

Table 3: Computational time for different algorithms for one mixture with $M = 2$ and $N = 3$

BSS-PALM	BSS-LPADMM	BSS-Deco	Demix-ℓ_1
1430 s	4780 s	2840 s	1280 s

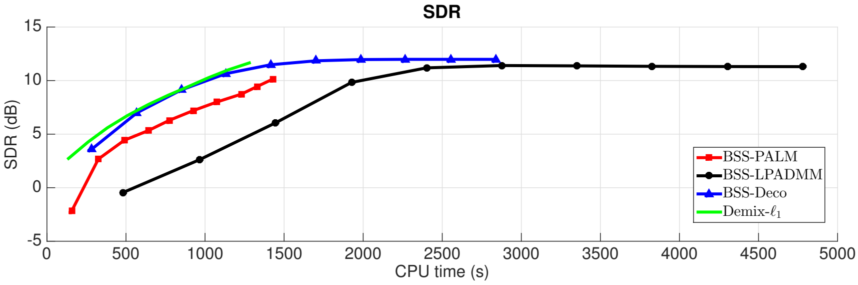

Figure 8: Evolution of the SDR with respect to the CPU time

6. Discussion and conclusion

In this paper, we studied the link between some ICA methods (FastICA and Infomax) and SCA for BSS with instantaneous mixtures. By combining the decorrelation constraint in time domain and the synthesis sparsity optimization, we proposed a new framework of SICA to generalize ICA into noisy and underdetermined scenario. We have shown that FastICA and Infomax can be formulated as a particular case of the proposed SICA functionnal (18). We designed several iterative algorithms to solve the problem.

Numerical experiments clearly support that taking the decorrelation constraint into account greatly improves the separation results in terms of SIR, without degrading the SDR. This claim is particularly supported by the fact that the proposed algorithms BSS-LPADMM and BSS-Deco outperform the oracle source separation algorithm without any decorrelation constraint. Moreover, except when the number of unknown sources is large, the proposed BSS-LPADMM and BSS-Deco reach comparable results of their oracle source separation with decorrelation constraint BSS-LPADMM-Oracle.

Regarding the computational time of the various approaches used in the experiments, it appears that the ICA approaches remain the most competitive for noiseless and overdetermined mixtures. However, for noisy overdetermined mixtures, BSS-Deco appears to be much more robust than GMCA with respect to the input SNR and the number of unknown sources, GMCA being already known to be more robust to noise than ICA methods [17]. Finally, for underdetermined mixtures (with or without noise) BSS-Deco appears to be very competitive: while its computational cost is twice that of Demix-ℓ_1, the SIR improvement is around 1 dB for an input SNR of 10 dB, and 2 dB for an input SNR of 20 dB, while the SDR is slightly higher (less than 1 dB).

Future work will focus on extending the SICA framework to convolutive mixtures. The straightforward extension of this work could also be considered: studying other sparse regularization than the simple ℓ_1 norm, such as social sparsity [56], $\frac{\ell_p}{\ell_q}$ criterion [57], but also considering the sparsity constraint directly on the analysis coefficient of the sources as in [58].

Appendix A.

Appendix A.1. Derivation of the PALM algorithm Alg. 1

PALM method is designed to deal with non-convex problems reading:

$$\min_{\mathbf{x},\mathbf{y}} H(\mathbf{x}) + Q(\mathbf{x},\mathbf{y}) + G(\mathbf{y}), \tag{A.1}$$

where $H(\mathbf{x})$ and $G(\mathbf{y})$ are proper lower semi-continuous functions, $Q(\mathbf{x},\mathbf{y})$ is a smooth function with Lipschitz gradient on any bounded set. The proximal method proposed in [22] updates the estimate of $(\mathbf{x},\mathbf{y})$ via

$$\mathbf{x}^{(j+1)} \in \operatorname*{argmin}_{\mathbf{x}} H(\mathbf{x}) + \langle \mathbf{x} - \mathbf{x}^{(j)}, \nabla_{\mathbf{x}} Q(\mathbf{x}^{(j)}, \mathbf{y}^{(j)}) \rangle + \frac{t^{1,(j)}}{2} \|\mathbf{x} - \mathbf{x}^{(j)}\|_2^2, \tag{A.2}$$

$$\mathbf{y}^{(j+1)} \in \operatorname*{argmin}_{\mathbf{y}} G(\mathbf{y}) + \langle \mathbf{y} - \mathbf{y}^{(j)}, \nabla_{\mathbf{y}} Q(\mathbf{x}^{(j+1)}, \mathbf{y}^{(j)}) \rangle + \frac{t^{2,(j)}}{2} \|\mathbf{y} - \mathbf{y}^{(j)}\|_2^2, \tag{A.3}$$

where $t^{1,(j)}$ and $t^{2,(j)}$ are two appropriate chosen step sizes. Thanks to the proximal operator

$$\operatorname{prox}_{\psi}(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2 + \psi(\mathbf{x}), \tag{A.4}$$

the minimization steps (A.2) and (A.3) can be written as follows:

$$\mathbf{x}^{(j+1)} \in \operatorname{prox}_{H/t^{1,(j)}} \left(\mathbf{x}^{(j)} - \frac{1}{t^{1,(j)}} \nabla_{\mathbf{x}} Q(\mathbf{x}^{(j)}, \mathbf{y}^{(j)}) \right), \tag{A.5}$$

$$\mathbf{y}^{(j+1)} \in \operatorname{prox}_{G/t^{2,(j)}} \left(\mathbf{y}^{(j)} - \frac{1}{t^{2,(j)}} \nabla_{\mathbf{y}} Q(\mathbf{x}^{(j+1)}, \mathbf{y}^{(j)}) \right). \tag{A.6}$$

It is shown in [22] that the sequence generated by (A.5) (A.6) converges to a critical point of (A.1). For problem (7), we have:

$$\begin{aligned} &H(\boldsymbol{\alpha}) = \lambda\|\boldsymbol{\alpha}\|_1 \ , \ G(\mathbf{A}) = \imath_{\mathcal{C}}(\mathbf{A}), \\ &Q(\boldsymbol{\alpha}, \mathbf{A}) = \frac{1}{2}\|\mathbf{X} - \mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2, \end{aligned} \tag{A.7}$$

and the particular choices:

$$t^{1,(j)} = L^{1,(j)}, \quad t^{2,(j)} = L^{2,(j)}, \tag{A.8}$$

where $L^{1,(j)}$ and $L^{2,(j)}$ are the Lipschitz constants of $\nabla_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}^{(j)}, \mathbf{A}^{(j)})$ and $\nabla_{\mathbf{A}} Q(\boldsymbol{\alpha}^{(j+1)}, \mathbf{A}^{(j)})$ respectively.

Knowing that the Proximal operator of ℓ_1 is the soft thresholding operator (20), and the proximal operator of $\imath_{\mathcal{C}}(\mathbf{A})$ is given by the normalization (21), one recovers Alg. 1.

Appendix A.2. Derivation of BSS-LPADMM Alg. 2

The augmented Lagrangian reads:

$$\begin{aligned} L(\mathbf{A},\boldsymbol{\alpha},\mathbf{S},\boldsymbol{\eta}) =& \frac{1}{2}||\mathbf{X}-\mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*||_F^2 + \lambda||\boldsymbol{\alpha}||_1 + \imath_{\mathcal{C}}(\mathbf{A}) \\ &+ \imath_{\text{deco}}(\mathbf{S}) + \langle \boldsymbol{\eta}, \mathbf{S}-\boldsymbol{\alpha}\boldsymbol{\Phi}^*\rangle + \frac{\gamma}{2}\|\mathbf{S}-\boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2, \end{aligned} \tag{A.9}$$

where $\boldsymbol{\eta}$ is the dual variable and γ is the penalty parameter. Let

$$F(\mathbf{A},\boldsymbol{\alpha},\mathbf{S},\boldsymbol{\eta}) = \frac{1}{2}||\mathbf{X}-\mathbf{A}\boldsymbol{\alpha}\boldsymbol{\Phi}^*||_F^2 + \langle \boldsymbol{\eta}, \mathbf{S}-\boldsymbol{\alpha}\boldsymbol{\Phi}^*\rangle + \frac{\gamma}{2}\|\mathbf{S}-\boldsymbol{\alpha}\boldsymbol{\Phi}^*\|_F^2. \tag{A.10}$$

LPADMM minimizes the augmented Lagrangian by iteratively updating the primal and dual variables via the following update rules:

$$\boldsymbol{\alpha}^{(j+1)} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\langle \nabla_{\boldsymbol{\alpha}} F(\mathbf{A}^{(j)},\boldsymbol{\alpha}^{(j)},\mathbf{S}^{(j)},\boldsymbol{\eta}^{(j)}),\boldsymbol{\alpha}\rangle + \lambda\|\boldsymbol{\alpha}\|_1 + \frac{L^{F,(j)}}{2}\|\boldsymbol{\alpha}^{(j)}-\boldsymbol{\alpha}\|_F^2, \tag{A.11}$$

$$\mathbf{S}^{(j+1)} = \underset{\mathbf{S}}{\operatorname{argmin}}\langle \boldsymbol{\eta}^{(j)}, \mathbf{S}-\boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*\rangle + \frac{\gamma}{2}\|\mathbf{S}-\boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*\|_F^2 + \imath_{\text{deco}}(\mathbf{S}), \tag{A.12}$$

$$\mathbf{A}^{(j+1)} = \underset{\mathbf{A}}{\operatorname{argmin}}\ \frac{1}{2}\|\mathbf{X}-\mathbf{A}\boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*\|_F^2 + \imath_{\mathcal{C}}(\mathbf{A}), \tag{A.13}$$

$$\boldsymbol{\eta}^{(j+1)} = \boldsymbol{\eta}^{(j)} + \gamma(\mathbf{S}^{(j+1)}-\boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^*). \tag{A.14}$$

In the sub-problem (A.11), $L^{F,(j)}$ is the Lipschitz constant of $\nabla_{\boldsymbol{\alpha}} F(\mathbf{A}^{(j)},\boldsymbol{\alpha}^{(j)},\mathbf{S}^{(j)},\boldsymbol{\eta}^{(j)})$, with

$$\nabla_{\boldsymbol{\alpha}} F(\mathbf{A}^{(j)},\boldsymbol{\alpha}^{(j)},\mathbf{S}^{(j)},\boldsymbol{\eta}^{(j)}) = -\mathbf{A}^{(j)^T}(\mathbf{X}-\mathbf{A}^{(j)}\boldsymbol{\alpha}^{(j)}\boldsymbol{\Phi}^*)\boldsymbol{\Phi} - \boldsymbol{\eta}^{(j)}\boldsymbol{\Phi} - \gamma(\mathbf{S}^{(j)}-\boldsymbol{\alpha}^{(j)}\boldsymbol{\Phi}^*)\boldsymbol{\Phi}. \tag{A.15}$$

Using the soft-thresholding operator (20), (A.11) can be rewritten as:

$$\boldsymbol{\alpha}^{(j+1)} = \mathcal{S}_{\lambda/L^{F,(j)}}\left(\boldsymbol{\alpha}^{(j)} - \frac{\nabla_{\boldsymbol{\alpha}} F(\mathbf{A}^{(j)},\boldsymbol{\alpha}^{(j)},\mathbf{S}^{(j)},\boldsymbol{\eta}^{(j)})}{L^{F,(j)}}\right). \tag{A.16}$$

The sub-problem (A.12) can be formulated as a decorrelation projection:

$$\begin{cases} \mathbf{S}^{(j+1)} = \underset{\mathbf{S}}{\operatorname{argmin}}\ \dfrac{\gamma}{2}\left\|\mathbf{S}-\boldsymbol{\alpha}^{(j+1)}\boldsymbol{\Phi}^* + \dfrac{\boldsymbol{\eta}^{(j)}}{\gamma}\right\|_F^2 \\ \text{s.t. } \mathbf{S}\mathbf{S}^T = \mathbf{D}, \text{ with } \mathbf{D} \text{ diagonal.} \end{cases} \tag{A.17}$$

and can be solved thanks to the following proposition [59]:

Proposition 2. *Let* $\mathbf{s}(t) \in \mathbb{R}^N$ *be a 0-mean random vector with a variance-covariance matrix* $\Sigma_{\mathbf{s}}$. *Let* $\mathbf{W}$ *be the optimal decorrelation transform that minimizes the Mean-Squared Error (MSE) between the input* $\mathbf{s}(t)$ *and the output* $\mathbf{y}(t) = \mathbf{W}\mathbf{s}(t)$, *such that its* *covariance matrix $\Sigma_{\mathbf{Y}}$ is diagonal:*

$$\min_{\mathbf{W}} \mathbb{E}\left\{(\mathbf{s}(t)-\mathbf{y}(t))^2\right\} \quad s.t.\ \mathbf{y}(t)=\mathbf{W}\mathbf{s}(t),\ \Sigma_{\mathbf{y}}=\sigma_{\mathbf{y}}^2\mathbf{I},$$

then $\mathbf{W}=\sigma_{\mathbf{y}}^2\Sigma_{\mathbf{s}}^{-1/2}$.

For the matrix of sources $\mathbf{S}$, when T is large, the empirical covariance matrix $\mathbf{S}\mathbf{S}^T$ converges to the covariance matrix $\Sigma_{\mathbf{s}}$. We can then choose in practice $\mathbf{W}=(\text{diag}(\Sigma_{\mathbf{S}}))^{1/2}\Sigma_{\mathbf{S}}^{-1/2}$.

Finally, the sub-problem (A.13) is tackled by a classical projected gradient descend.

[1] S. Kim, C. D. Yoo, Underdetermined blind source separation based on subspace representation, IEEE Transactions on Signal processing 57 (7) (2009) 2604–2614.

[2] S.-i. Amari, A. Cichocki, Adaptive blind signal processing-neural network approaches, Proceedings of the IEEE 86 (10) (1998) 2026–2048.

[3] A. Mansour, A. K. Barros, N. Ohnishi, Blind separation of sources: Methods, assumptions and applications, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences 83 (8) (2000) 1498–1512.

[4] J. Hérault, B. Ans, Réseau de neurones à synapses modifiables: Décodage de messages sensoriels composites par apprentissage non supervisé et permanent, Comptes rendus des séances de l'Académie des sciences. Série 3, Sciences de la vie 299 (13) (1984) 525–528.

[5] P. Comon, Independent component analysis, a new concept?, Signal processing 36 (3) (1994) 287–314.

[6] P. Comon, C. Jutten, Handbook of Blind Source Separation: Independent component analysis and applications, Academic press, 2010.

[7] R. Gribonval, S. Lesage, A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges, in: Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN), d-side publi., 2006, pp. 323–330.

[8] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, T. J. Sejnowski, Removing electroencephalographic artifacts by blind source separation, Psychophysiology 37 (02) (2000) 163–178.

[9] M. J. McKeown, T. J. Sejnowski, et al., Independent component analysis of fMRI data: examining the assumptions, Human brain mapping 6 (5-6) (1998) 368–372.

[10] R. Ge, Y. Wang, J. Zhang, L. Yao, H. Zhang, Z. Long, Improved FastICA algorithm in fMRI data analysis using the sparsity property of the sources, Journal of neuroscience methods 263 (2016) 103–114.

[11] M. Plumbley, S. Abdallah, J. Bello, M. Davies, J. Klingseisen, G. Monti, M. Sandler, ICA and related models applied to audio analysis and separation, in: Proceedings of the 4th International ICSC Symposium on Soft Computing and Intelligent Systems for Industry, Citeseer, 2001.

[12] P. Bofill, M. Zibulevsky, Underdetermined blind source separation using sparse representations, Signal processing 81 (11) (2001) 2353–2362.

[13] M. Babaie-Zadeh, C. Jutten, A. Mansour, Sparse ICA via cluster-wise PCA, Neurocomputing 69 (13) (2006) 1458–1466.

[14] S. Arberet, R. Gribonval, F. Bimbot, A robust method to count and locate audio sources in a multichannel underdetermined mixture, IEEE Transactions on Signal Processing 58 (1) (2010) 121–133.

[15] O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE transactions on Signal Processing 52 (7) (2004) 1830–1847.

[16] Y. Li, S.-I. Amari, A. Cichocki, D. W. Ho, S. Xie, Underdetermined blind source separation based on sparse representation, IEEE Transactions on Signal Processing 54 (2) (2006) 423–437.

[17] J. Bobin, J.-L. Starck, Y. Moudden, M. J. Fadili, Blind source separation: the sparsity revolution, Advances in Imaging and Electron Physics 152 (2008) 221–302.

[18] H. Zayyani, M. Babaie-Zadeh, C. Jutten, An iterative bayesian algorithm for sparse component analysis in presence of noise, IEEE Transactions on Signal Processing 57 (11) (2009) 4378–4390.

[19] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D'ardenne, W. Richter, J. Cohen, J. Haxby, Independent component analysis for brain fMRI does not select for independence, Proceedings of the National Academy of Sciences 106 (26) (2009) 10415–10422.

[20] V. D. Calhoun, V. K. Potluru, R. Phlypo, R. F. Silva, B. A. Pearlmutter, A. Caprihan, S. M. Plis, T. Adalı, Independent component analysis for brain fMRI does indeed select for maximal independence, PloS one 8 (8) (2013) e73309.

[21] F. Feng, M. Kowalski, A unified approach for blind source separation using sparsity and decorrelation, in: Proceedings of the 23nd European Signal Processing Conference (EUSIPCO), IEEE, 2015, pp. 1736–1740.

[22] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Mathematical Programming (2013) 1–36.

[23] M. Davies, Audio source separation, in: Institute of mathematics and its applications conference series, Vol. 71, Oxford; Clarendon; 1999, 2002, pp. 57–68.

[24] J. Hurri, A. Hyvärinen, E. Oja, Wavelets and natural image statistics, in: In Proc. Scandinavian Conf. on Image Analysis' 97, 1997, pp. 13–18.

[25] M. U. Khalid, A.-K. Seghouane, Multi-subject fMRI connectivity analysis using sparse dictionary learning and multiset canonical correlation analysis, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE, 2015, pp. 683–686.

[26] M. Zibulevsky, B. A. Pearlmutter, Blind source separation by sparse decomposition in a signal dictionary, Neural computation 13 (4) (2001) 863–882.

[27] X. Jia, M. Zhao, M. Buzza, Y. Di, J. Lee, A geometrical investigation on the generalized lp/lq norm for blind deconvolution, Signal Processing 134 (Supplement C) (2017) 63 – 69.

[28] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non-gaussian signals, in: IEE Proceedings F (Radar and Signal Processing), Vol. 140, IET, 1993, pp. 362–370.

[29] L. De Lathauwer, J. Castaing, J.-F. Cardoso, Fourth-order cumulant-based blind identification of underdetermined mixtures, IEEE Transactions on Signal Processing 55 (6) (2007) 2965–2973.

[30] A. Belouchrani, A. Cichocki, Robust whitening procedure in blind source separation context, Electronics letters 36 (24) (2000) 2050–2051.

[31] M. Kowalski, A. Gramfort, Inverse problems with time-frequency dictionaries and non-white gaussian noise, in: Signal Processing Conference (EUSIPCO), 2015 23rd European, IEEE, 2015, pp. 1741–1745.

[32] F. Abrard, Y. Deville, A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources, Signal Processing 85 (7) (2005) 1389–1403.

[33] S. S. Chen, D. L. Donoho, M. A. Saunders, Atomic decomposition by basis pursuit, SIAM review 43 (1) (2001) 129–159.

[34] E. Vincent, Complex nonconvex ℓ_p norm minimization for underdetermined source separation, in: Proceedings of International Conference on Independent Component Analysis and Signal Separation (LVA/ICA), Springer, 2007, pp. 430–437.

[35] Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical programming 103 (1) (2005) 127–152.

[36] J. Mc, J.-L. Starck, J. Fadili, Y. Moudden, Sparsity and morphological diversity in blind source separation, IEEE Transactions on Image Processing 16 (11) (2007) 2662–2674.

[37] D. L. Donoho, A. G. Flesia, Can recent innovations in harmonic analysis "explain" key findings in natural image statistics?, Network: computation in neural systems 12 (3) (2001) 371–393.

[38] Y. Ouyang, Y. Chen, G. Lan, E. Pasiliao Jr, An accelerated linearized alternating direction method of multipliers, SIAM Journal on Imaging Sciences 8 (1) (2015) 644–681.

[39] A. Javaheri, H. Zayyani, F. Marvasti, Recovery of missing samples using sparse approximation via a convex similarity measure, in: Sampling Theory and Applications (SampTA), 2017 International Conference on, IEEE, 2017, pp. 543–547.

[40] R. Chartrand, B. Wohlberg, A nonconvex ADMM algorithm for group sparsity with sparse groups, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 6009–6013.

[41] Y. Wang, W. Yin, J. Zeng, Global convergence of ADMM in nonconvex nonsmooth optimization, arXiv preprint arXiv:1511.06324.

[42] L. Yang, T. K. Pong, X. Chen, Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction, arXiv preprint arXiv:1506.07029.

[43] Y. Li, A. Cichocki, S.-i. Amari, Analysis of sparse representation and blind source separation, Neural computation 16 (6) (2004) 1193–1234.

[44] F. Feng, M. Kowalski, Hybrid model and structured sparsity for under-determined convolutive audio source separation, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014.

[45] S. Araki, F. Nesta, E. Vincent, Z. Koldovskỳ, G. Nolte, A. Ziehe, A. Benichoux, The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation, in: Proceedings of Latent Variable Analysis and Signal Separation (LVA/ICA), Springer, 2012, pp. 414–422.

[46] P. L. Sondergaard, B. Torrésani, P. Balazs, The linear time frequency analysis toolbox, International Journal of Wavelets, Multiresolution and Information Processing 10 (04).

[47] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE Transactions on Audio, Speech, and Language Processing 14 (4) (2006) 1462–1469.

[48] R. Giryes, M. Elad, Y. C. Eldar, The projected GSURE for automatic parameter tuning in iterative shrinkage methods, Applied and Computational Harmonic Analysis 30 (3) (2011) 407–422.

[49] C.-A. Deledalle, S. Vaiter, J. Fadili, G. Peyré, Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection, SIAM Journal on Imaging Sciences 7 (4) (2014) 2448–2487.

[50] V. Emiya, E. Vincent, N. Harlander, V. Hohmann, Subjective and objective quality assessment of audio source separation, IEEE Transactions on Audio, Speech, and Language Processing 19 (7) (2011) 2046–2057.

[51] E. T. Hale, W. Yin, Y. Zhang, Fixed-point continuation for ℓ_1-minimization: Methodology and convergence, SIAM Journal on Optimization 19 (3) (2008) 1107–1130.

[52] Z. Koldovskỳ, P. Tichavskỳ, E. Oja, Efficient variant of algorithm fastica for independent component analysis attaining the Cramér-Rao lower bound, IEEE Transactions on neural networks 17 (5) (2006) 1265–1277.

[53] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, E. Moulines, A blind source separation technique using second-order statistics, IEEE Transactions on signal processing 45 (2) (1997) 434–444.

[54] A. Cichocki, S.-I. Amari, K. Siwek, T. Tanaka, A. H. Phan, R. Zdunek, S. Cruces, P. Georgiev, Y. Washizawa, Z. Leonowicz, ICALAB toolboxes, URL: http://www. bsp. brain. riken. jp/ICALAB.

[55] J. Bobin, J. Rapin, A. Larue, J.-L. Starck, Sparsity and adaptivity for the blind separation of partially correlated sources 63 (5) (2014) 1199–1213.

[56] M. Kowalski, K. Siedenburg, M. Dörfler, Social sparsity! neighborhood systems enrich structured shrinkage operators, IEEE transactions on signal processing 61 (10) (2013) 2498–2511.

[57] L. Li, Sparsity-promoted blind deconvolution of ground-penetrating radar (GPR) data, IEEE Geoscience and Remote Sensing Letters 11 (8) (2014) 1330–1334.

[58] S. Arberet, P. Vandergheynst, R. Carrillo, J. Thiran, Y. Wiaux, Sparse reverberant audio source separation via reweighted analysis, IEEE Transactions on Audio, Speech, and Language Processing 21 (7) (2013) 1391–1402.

[59] Y. C. Eldar, A. V. Oppenheim, MMSE whitening and subspace whitening, IEEE Transactions on Information Theory 49 (7) (2003) 1846–1851.