



HAL
open science

A Unified Approach for Over and Under-Determined Blind Source Separation Based on Both Sparsity and Decorrelation

Fangchen Feng, Matthieu Kowalski

► **To cite this version:**

Fangchen Feng, Matthieu Kowalski. A Unified Approach for Over and Under-Determined Blind Source Separation Based on Both Sparsity and Decorrelation. 2016. hal-01297471v1

HAL Id: hal-01297471

<https://hal.science/hal-01297471v1>

Preprint submitted on 4 Apr 2016 (v1), last revised 22 May 2018 (v3)

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

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

A Unified Approach for Over and Under-Determined Blind Source Separation Based on Both Sparsity and Decorrelation

Fangchen Feng, Matthieu Kowalski

Abstract—Over the last decades, independent component analysis (ICA) has been a major tool for blind source separation (BSS). Both theoretical and practical evaluations showed that the hypothesis of independence suits well for audio and musical signals. In the last few years, sparsity-based optimization has emerged as another efficient implement for BSS. This paper starts from introducing some new BSS methods that take advantage of both decorrelation, which is a direct consequence of independence, and sparsity using overcomplete Gabor representation. Theoretical proof and discussion supporting the convergence of the proposed algorithms are then presented. Numerical results are given illustrating the good performances of these approaches and their robustness to noise.

I. INTRODUCTION

In many situations, such as a concert for music or the so called cocktail party problem for speech, the recorded sound signals are issued from mixtures of several sound sources. The linear mixture model assumes that the problem can be formulated as follow :

$$x_m(t) = \sum_{n=1}^N A_{mn} s_n(t) + e_m(t) , \quad (1)$$

with N source signals s_n of duration T and M microphones, yielding M mixture channels x_m . The recording is supposed to be *instantaneous* and modeled by a set of mixing coefficients A_{mn} . Denoting by $\mathbf{x} \in \mathbb{R}^{M \times T}$ and $\mathbf{s} \in \mathbb{R}^{N \times T}$ the matrices of mixture channels and source signals respectively and by $\mathbf{A} \in \mathbb{R}^{M \times N}$ the mixing matrix, the process (1) can be rewritten more concisely in matrix form as :

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{e} , \quad (2)$$

where $\mathbf{e} \in \mathbb{R}^{M \times T}$ models the background noise. The goal of BSS is to reconstruct the sources \mathbf{s} and the mixing matrix \mathbf{A} only from the observations \mathbf{x} .

Most of the Blind Source Separation techniques can be divided into over-determined, determined and under-determined case according to the number of microphones and the number of sources. When the number of microphones is larger or equal to the number of sources (over-determined or determined scenario, respectively), the well-known Independent Component

Analysis (ICA) methods [25] are often applied with success. These methods assume that the sources are statistically independent and intend to achieve separation by minimizing an independence criterion between the components of the estimated sources. These methods have already provided successful results in a wide range of applications such as Electro-EncephaloGraphy (EEG) [27], functional Magnetic Resonance Imaging (fMRI) [29] and audio source separation [31].

Apart from the ICA, sparsity-based optimization methods have also been employed (see [17] and references therein). These methods first reformulate the problem in an optimization framework by using some sparsity properties of signals inside an appropriate dictionary of waveforms. Then some optimization algorithms such as iterative thresholding [14] can be used to solve the problem. These sparse approaches are mainly (but not only) used in the under-determined setting, *i.e.* when the number of microphones is less than the number of sources, in two steps: the mixing system is first estimated using clustering methods and the source signals are then estimated thanks to optimization approaches.

Recently, it has been claimed in [18] that some ICA methods (more specifically Infomax [10] and FastICA [26]) separate sparse sources rather than independent sources for fMRI in certain cases. It has been concluded that rather than decomposition methods that search for independent components, one should develop alternate methods targeting decompositions into components that are sparse. Starting from this remark, we construct in this article algorithms that benefit from both decorrelation, which is a direct consequence of independence, and sparsity. We show that these algorithms can be employed in both under-determined and over-determined cases and illustrate that by making such a combination, better separation performances can be achieved.

Extending [21], contributions of this article are threefold: first, we offer a combination of the decorrelation assumption and the sparse model. This is the first time, to our knowledge, that such a combination is used for BSS. Second, we propose several efficient optimization algorithms mostly with proven convergence which work in both under-determined and over-determined case. Finally, extensive experiments show the benefits of the proposed approach, as well as the advantages and limitations of popular BSS algorithms.

The rest of this paper is organized as follow. Section II provides an overview of the state of the art for BSS methods based on ICA and sparsity. In Section III, we present two separation methods based on alternating optimization and

F. Feng and M. Kowalski are with Laboratoire des Signaux et Systèmes, UMR 8506 CNRS - CENTRALESUPELEC - Univ Paris-Sud, 91192 Gif-sur-Yvette Cedex, France (e-mail: forname.name@l2s.centralesupelec.fr). This work benefited from the support of the FMJH Program Gaspard Monge in optimization and operation research, and from the support to this program from EDF

Section IV shows three proposed algorithms that combine decorrelation and sparsity. Finally, the numerical results are illustrated in Section V.

II. STATE OF THE ART

In this section, we give a brief introduction of the state of the art methods, mainly ICA and methods based on sparsity.

A. Independent Component Analysis

Techniques of ICA have been introduced at the beginning of 1980's by J.Herault, C.Jutten, and B.Ans [1], [23], [24], and are based on the statistical independence of the source signals modeled as random process [16]. Fundamentally, for the instantaneous linear mixing problem, ICA methods are designed to estimate a demix matrix $\mathbf{W} = \mathbf{A}^{-1}$ in order to obtain the estimated sources $\mathbf{y} = \mathbf{W}\mathbf{x}$, which are statistically independent. In [16], it has been demonstrated that the independence of the outputs leads to a matrix \mathbf{W} that satisfies $\mathbf{W}\mathbf{A} = \mathbf{P}\mathbf{D}$, where \mathbf{P} is a permutation matrix and \mathbf{D} is a diagonal matrix, which means that the estimated outputs and the sources signals are equal up to a permutation and a scale factor.

However, the penalty of independence demands not only the equality of two multivariate functions (the probability density function) but also the perfect knowledge of them. Therefore, several independence measurements have been proposed. Fundamentally, central limit theorem tells us that mixing independent signals should lead to a kind of Gaussianization [6]. Then, it seems natural that demixing leads to process that deviate from Gaussian process. Actually, most of the ICA methods maximize the non-gaussianity of the estimated signals to perform separation. However, at the same time, ICA methods based on Information Maximization, Maximum Likelihood and Higher Order Statistics have been developed over the years (see [6] and references therein). Interestingly, as pointed out in [28], most ICA-based algorithms have similar properties and results.

Despite its theoretical strength and excellent experimental results, the hypothesis of ICA brings some limits to the model as well. First, "independence is nongaussianity" assumption brings the constraint that the independent components should have the non-gaussian distribution. Second, as mentioned in [6], ICA algorithms require information on the sources distribution since the way the demix matrix (and thus the sources) is estimated closely depends on the model chosen on the sources. Third, because of the linear recovery, ICA methods work only in determined case. At last, as mentioned in [19], noise clearly degenerates the ICA model and makes the sources not fully identifiable.

Recently, it was claimed in [18] that two of the most used ICA methods for fMRI (Infomax and FastICA) separate sparse sources rather than independent sources, leading to the conclusion that the mathematical design of better analysis tool for brain fMRI should emphasize on other characteristics, such as sparsity, rather than independence. One given explanation is that the classic 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 the generalized Gaussian distribution, which includes the sparse sources modeled by ℓ_1 minimization method. Although the experiments made in [18] are only concentrated on InfoMax and FastICA (not the ICA methods in general) and the conclusion is only drawn for the brain fMRI sources, this work still motivates us to pay more attention to sparsity-based methods and to bring some independence assumptions into the designed algorithms.

B. Sparsity and GMCA method

The seminal paper [37] has introduced a BSS method that focuses on sparsity to distinguish the sources. The sources are assumed to be sparse in a particular basis \mathcal{D} . Then, the sources \mathbf{s} and the mixing matrix \mathbf{A} are estimated from a maximum a posteriori estimator of the sources in the basis \mathcal{D} . They showed that sparsity clearly enhances the diversity between the sources.

1) *Sparsity in the over-determined case:* More precisely, as claimed in [5], sparsity based methods have several advantages. In one hand, the sparser the sources in the dictionary, the more "separable" they are. On the other hand, sparse sources being concentrated on few significant coefficients in the sparse domain, the noise appears as a slight perturbation. Thus, sparsity based methods should be less sensitive to noise. Based on these remarks, they proposed the Generalized Morphological Component Analysis (GMCA). First, they define \mathcal{D} as the concatenation of D orthonormal bases $\{\Phi_i\}_{i=1, \dots, D} : \mathcal{D} = [\Phi_1^T, \dots, \Phi_D^T]$, then they assume *a priori* that the sources s_n are sparse in the dictionary \mathcal{D} . In this GMCA setting, each source is modeled as a linear combination of D morphological components:

$$s_n = \sum_{k=1}^D \varphi_{nk} = \sum_{k=1}^D \alpha_{nk} \Phi_k \quad (3)$$

Then, an optimization formulation for the estimation of \mathbf{A} and \mathbf{s} is proposed:

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{s}}\} = \underset{\mathbf{A}, \mathbf{s}}{\operatorname{argmin}} \sum_{n=1}^N \sum_{k=1}^D \|\varphi_{nk} \Phi_k^T\|_1 + \kappa \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_2^2 \quad (4)$$

In order to minimize (4), the authors use a classical alternating optimization techniques: the sources are estimated for a fixed mixing system, and then mixing system is estimated to fixed sources. This GMCA method has shown good performances, especially when the signals are corrupted by Gaussian white noise.

2) *The under-determined case:* Most of the methods in the (over-)determined case do not work directly in the under-determined scenario. In this case, two-steps methods are often used. Based on the sparsity assumption, the mixing matrix is first estimated and the estimation of the sources is performed supposing that the mixing matrix is known.

Thanks to the sparsity of the sources, the mixtures tend to cluster along the mixing matrix columns. Thus, estimating the mixing matrix amounts to identify linear clusters passing through the origin (we suppose here that the sources have zero-mean) and then merely defined by their angle to the origin.

Once the mixture is determined, estimating the sources can be done by typical ℓ_p optimization approaches. In practice, when the mixing system \mathbf{A} is known, estimation of the sources reduces to the minimization of (4) for a fixed \mathbf{A} .

As a matter of fact, there exists various methods based on time-frequency (T-F) masking or maximum a posteriori (MAP) (see [32] and references therein). Here we take the DUET method [35] as reference. This method is based on the assumption that the supports of STFT of the sources are disjoint. First, they define a power weighted two-dimensional histogram constructed from the ratio of the time-frequency representations of the mixtures that is shown to have one peak for each source with peak location corresponding to the relative attenuation and delay mixing parameters. Then, this histogram is used to create time-frequency masks that partition one of the mixtures into the original sources.

This method is very fast and easy to compute. However, the assumption that the time-frequency representations of the source signals do not overlap is often too strong for audio signals especially when the number of sources is large. Besides, it has the common disadvantage of two-step methods: the estimation error of the mixing matrix will sometimes bring negative influence in the procedure of source estimation.

In the following, we will focus on sparsity-based optimization and try to bring the decorrelation assumption into the procedure to design algorithms that work in both under-determined and over-determined case. In the next Section, we give a convergence study of an intuitive algorithm based on GMCA which applies the decorrelation implicitly and we also propose a new method based on alternating optimization.

III. GMCA-LIKE METHOD AND PROXIMAL ALTERNATING LINEARIZED MINIMIZATION

In this section, we give a quick review of a classic separation method based only on sparsity which is a direct derivative from the GMCA method. We also give a convergence study of this method and show that the convergence counts on the decorrelation assumption.

A. GMCA-like method

In the setting of GMCA, the authors assume that the signals are sparse in some orthonormal bases. However, for audio signals, the overcomplete Gabor frame is often used as the dictionary. Then, we take here the so-called *synthesis* setting instead of the *analysis* point of view as in (4).

Let us denote by $\Phi \in \mathbb{C}^{T \times B}$ the matrix representing an energy-preserving STFT operator (or Parseval Gabor frame), the sources \mathbf{s} can be resynthesized from their estimated STFT coefficients $\alpha \in \mathbb{C}^{N \times B}$ by

$$\mathbf{s} = \alpha \Phi^* \quad (5)$$

where $\Phi^* \in \mathbb{C}^{B \times T}$ is the adjoint operator of Φ , that is its Hermitian transpose.

We use the following classic optimization framework to model the separation problem:

$$\min_{\alpha, \mathbf{A}} f(\mathbf{x}, \mathbf{A}, \alpha) + \Psi(\alpha) + g(\mathbf{A}) \quad (6)$$

where the data term f measures the fit between the observed mixture \mathbf{x} and the sources synthesis Gabor coefficients α via the mixing matrix \mathbf{A} . We stick here to a classical ℓ_2 data term corresponding to a Gaussian prior on the noise in a Bayesian point of view

$$f(\mathbf{x}, \mathbf{A}, \alpha) = \frac{1}{2} \|\mathbf{x} - \mathbf{A}\alpha\Phi^*\|_F^2 .$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The classical ℓ_1 regularization term is employed in order to favor sparse solution

$$\Psi(\alpha) = \|\alpha\|_1 .$$

It should be mentioned that various other choices for the penalty term can be made (see for example [3] and references therein), but we stick here to the ℓ_1 norm for the sake of simplicity. Finally, $g(\mathbf{A})$ is mainly used to avoid the separation ambiguity [8] such as the scaling and permutation problem between \mathbf{s} and \mathbf{A} . In order to obtain a convex functional with respect to \mathbf{A} , α being fixed, we can set $g(\mathbf{A}) = \nu_B$, *i.e.* the indicator function of the unit ball

$$\nu_B(\mathbf{A}) = \begin{cases} 0 & \text{if } \|\mathbf{a}_n\| \leq 1, n = 1, 2, \dots, N \\ +\infty & \text{otherwise} \end{cases} \quad (7)$$

where \mathbf{a}_n is the n -th column of \mathbf{A} . We then obtain the classical ℓ_2 - ℓ_1 problem [36].

$$\min_{\alpha, \mathbf{A}} \frac{1}{2} \|\mathbf{x} - \mathbf{A}\alpha\Phi^*\|_2^2 + \lambda \|\alpha\|_1 + \nu_B(\mathbf{A}) \quad (8)$$

Now, one apply the idea of GMCA to solve the problem (8). First, the source signals are estimated considering that the mixing system is fixed:

$$\tilde{\alpha} = \arg \min_{\alpha} \frac{1}{2} \|\mathbf{x} - \mathbf{A}\alpha\Phi^*\|_F^2 + \lambda \|\alpha\|_1 . \quad (9)$$

Then, the actualization of the mixing system using the estimated sources is a least squares problem:

$$\begin{aligned} \tilde{\mathbf{A}} &= \min_{\mathbf{A}} \frac{1}{2} \|\mathbf{x} - \mathbf{A}\alpha\Phi^*\|_F^2 \\ &= \mathbf{x}\Phi\alpha^*(\alpha\Phi^*\Phi\alpha^*)^{-1} \end{aligned} \quad (10)$$

At last, the normalization on each columns of the matrix \mathbf{A} is performed to avoid the separation ambiguity.

Φ being an overcomplete dictionary, the sub-problem (9) can be solved efficiently with iterative algorithms, such as proximal algorithms [4], [14]. However, despite an excellent behavior in practice, such an alternating algorithm does not benefit of any convergence guaranty.

B. Two convergent algorithms

Thanks to the recent works [7], [12], we propose two convergent procedures. The notion of proximity operator is reminded, as the two algorithms rely on it.

Definition 1. Let $\psi : \mathbb{C}^{M \times N} \rightarrow (-\infty, +\infty]$ be a proper, lower semicontinuous function, let $\mathbf{U} \in \mathbb{R}^{N \times N}$ by a symmetric positive definite matrix, and let $\mathbf{x} \in \mathbb{C}^{M \times N}$. We denote by *The*

proximity operator of ψ at \mathbf{x} relative to the matrix induced by \mathbf{U} is defined as :

$$\text{prox}_{\mathbf{U},\psi} := \underset{\mathbf{y} \in \mathbb{C}^{M \times N}}{\text{argmin}} \psi(\mathbf{y}) + \frac{1}{2} \|\mathbf{y} - \mathbf{x}\|_{\mathbf{U}}^2$$

where $\|\mathbf{x}\|_{\mathbf{U}} = \sqrt{\langle \mathbf{x}\mathbf{U}, \mathbf{x} \rangle}$.

We note $\text{prox}_{\mathbf{I},\psi} = \text{prox}_{\psi}$.

The first procedure is the well known Proximal Alternating Linearized Minimization (PALM) algorithm carefully studied in [7] in the non convex setting. This procedure has the great advantage to be very simple to implement and is given in Algorithm 1. Here, prox_g is simply the projection on the unit ball, *i.e.* consists of the normalization of the columns of the matrix.

Algorithm 1: PALM

Initialization : $\alpha_1 \in C^{N \times B}$, $\mathbf{A}_1 \in R^{M \times N}$,
 $L_{1,1} = \|\mathbf{A}_1\|_F^2$, $L_{1,2} = \|\alpha_1 \Phi^*\|_F^2$, $k = 1$,

repeat

$$\begin{aligned} \nabla_{\alpha} f(\mathbf{A}_k, \alpha_k) &= -\mathbf{A}_k^T (\mathbf{x} - \mathbf{A}_k \alpha \Phi^*) \Phi; \\ \alpha_{k+1} &= \text{prox}_{\Psi/L_{k,1}} \left(\alpha_k - \frac{1}{L_{k,1}} \nabla_{\alpha} f(\mathbf{A}_k, \alpha_k) \right); \\ \nabla_{\mathbf{A}} f(\mathbf{A}_k, \alpha_{k+1}) &= -(\mathbf{x} - \mathbf{A} \alpha_{k+1} \Phi^*) \Phi \alpha_{k+1}^*; \\ \mathbf{A}_{k+1} &= \text{prox}_{g/L_{k,2}} \left(\mathbf{A}_k - \frac{1}{L_{k,2}} \nabla_{\mathbf{A}} f(\mathbf{A}_k, \alpha_{k+1}) \right); \\ L_{k+1,1} &= \|\mathbf{A}_{k+1}\|_F^2; \\ L_{k+1,2} &= \|\alpha_{k+1} \Phi^*\|_F^2; \\ k &= k + 1; \end{aligned}$$

until convergence;

The second approach is the The Block Coordinate Variable Metric Forward-Backward algorithm (BC-VMFB) [12]. This last approach, given in Algorithm 2, is more difficult to implement, but will illustrate the fact that enforcing the decorrelation can be helpful.

Algorithm 2: BC-VMFB

Initialization : $\alpha_1 \in C^{N \times B}$, $\mathbf{A}_1 \in R^{M \times N}$,
 $L_{1,1} = \|\mathbf{A}_1\|_F^2$, $k = 1$;

repeat

$$\begin{aligned} \nabla_{\alpha} f(\mathbf{A}_k, \alpha_k) &= -\mathbf{A}_k^T (\mathbf{x} - \mathbf{A}_k \alpha \Phi^*) \Phi; \\ \alpha_{k+1} &= \text{prox}_{\Psi/L_{k,1}} \left(\alpha_k - \frac{1}{L_{k,1}} \nabla_{\alpha} f(\mathbf{A}_k, \alpha_k) \right); \\ \nabla_{\mathbf{A}} f(\mathbf{A}_k, \alpha_{k+1}) &= -(\mathbf{x} - \mathbf{A} \alpha_{k+1} \Phi^*) \Phi \alpha_{k+1}^*; \\ \mathbf{H}_k &= \alpha_{k+1} \Phi^* \Phi \alpha_{k+1}^*; \\ \mathbf{A}_{k+1} &= \text{prox}_{\mathbf{H}_k, g} \left(\mathbf{A}_k - \nabla_{\mathbf{A}} f(\mathbf{A}_k, \alpha_{k+1}) \mathbf{H}_k^{-1} \right); \\ L_{k+1,1} &= \|\mathbf{A}_{k+1}\|_F^2; \\ k &= k + 1; \end{aligned}$$

until convergence;

The first proximal descent with respect to α_k is the same as in PALM. The main difference is during the estimation of the mixing matrix \mathbf{A} , where the proximal descent is preconditioned by the second order derivative of f with respect to \mathbf{A} :

$$\mathbf{H}_k = \frac{\partial f(\mathbf{A}, \alpha_{k+1})^2}{\partial^2 \mathbf{A}} = \alpha_{k+1} \Phi^* \Phi \alpha_{k+1}^* .$$

Then, this proximal step becomes

$$\begin{aligned} \mathbf{A}_{k+1/2} &= \mathbf{x} \Phi \alpha_{k+1}^* (\alpha_{k+1} \Phi^* \Phi \alpha_{k+1}^*)^{-1} \\ \mathbf{A}_{k+1} &= \mathcal{P}_{H_k, \iota_B}(\mathbf{A}_{k+1/2}) \end{aligned} \quad (11)$$

Interestingly enough, one can notice that the least square estimate of \mathbf{A} is projected on the unit ball *relatively* to the matrix \mathbf{H}_k . If this projection is hard in general, it becomes very simple if in every iteration the estimated sources are decorrelated. Indeed, in that case the proximal step simply becomes $\mathbf{A}_{k+1} = \mathbf{x} \Phi \alpha_{k+1}^*$ followed by the projection on the unit ball.

For now, we do not use any strategy to force decorrelation, so in the next section, we will present several methods that combine sparsity and decorrelation.

IV. BSS WITH SPARSITY AND DECORRELATION

In this section, we present three methods to consider the decorrelation constraint. We first show that the decorrelation can be obtained by simply adding a regularization term based on PALM method. We then propose a functional taking this constraint into account, minimized thanks to a method based on ADMM. At last we present one empirical method based on the study of GMCA-like method which is easier to implement.

Supposing that \mathbf{s} has zero mean, decorrelation assumes that

$$\mathbb{E}(\mathbf{s}\mathbf{s}^T) = \mathbf{D} , \quad (12)$$

where \mathbf{D} is a diagonal matrix and \mathbb{E} is the expectation operator. Numerically, as the number of samples is large, we simply assume that

$$\mathbf{s}\mathbf{s}^T = \mathbf{D} . \quad (13)$$

This decorrelation assumption differs from the whitening procedure by the fact that elements on the diagonal matrix \mathbf{D} are unknown.

We start by showing that this numerical decorrelation can be enforced by adding a regularization term based on the GMCA-like method.

A. Regularized GMCA

Rather than (8), we propose to minimize the following functional:

$$\min_{\alpha, \mathbf{A}} \frac{1}{2} \|\mathbf{x} - \mathbf{A} \alpha \Phi^*\|_2^2 + \frac{\mu}{2} \|\mathbf{A}\|_2^2 + \lambda \|\alpha\|_1 + \iota_{\mathcal{C}}(\mathbf{A}) \quad (14)$$

where $\iota_{\mathcal{C}}(\mathbf{A})$ is the indicator function of the unit circle

$$\iota_{\mathcal{C}}(\mathbf{A}) = \begin{cases} 0 & \text{if } \|\mathbf{a}_n\| = 1, n = 1, 2, \dots, N \\ +\infty & \text{otherwise} \end{cases} \quad (15)$$

Compared to GMCA-like methods, a ℓ_2 regularizer is added on the mixing matrix \mathbf{A} , and the constraint on the ℓ_2 ball is replaced by the ℓ_2 circle.

The BC-VMFB algorithm is applied and we choose $f(\mathbf{A}, \alpha) = \frac{1}{2} \|\mathbf{x} - \mathbf{A} \alpha \Phi^*\|_2^2 + \frac{\mu}{2} \|\mathbf{A}\|_2^2$ in Alg. 2. With the particular choice

$$\mathbf{G}_k = L_k \text{ and } \mathbf{H}_k = \frac{\partial f(\mathbf{A}, \alpha_{k+1})^2}{\partial^2 \mathbf{A}} = \alpha_{k+1} \Phi^* \Phi \alpha_{k+1}^* + \mu \mathbf{I} ,$$

the proximal step for \mathbf{A} becomes:

$$\begin{aligned} \mathbf{A}_{k+1/2} &= \mathbf{x}\Phi\alpha_{k+1}^* (\alpha_{k+1}\Phi^*\Phi\alpha_{k+1}^* + \mu\mathbf{I})^{-1} \\ \mathbf{A}_{k+1} &\in \text{prox}_{\mathbf{H}_k, \iota_C}(\mathbf{A}_{k+1/2}) \end{aligned} \quad (16)$$

Because of the constraint on the unit circle, the choice of the parameter μ does not change the minimizer of (14). Then, by choosing μ large enough in order to have

$$(\alpha_{k+1}\Phi^*\Phi\alpha_{k+1}^* + \mu\mathbf{I})^{-1} = \frac{1}{\mu}\mathbf{I} + \varepsilon$$

where ε is of order of the machine precision, the projection simply reads:

$$\begin{aligned} \mathbf{A}_{k+1/2} &= \mathbf{x}\Phi\alpha_{k+1}^*/\mu \\ \mathbf{a}_{k+1,n} &= \mathbf{a}_{k+1/2,n}/\|\mathbf{a}_{k+1/2,n}\|, \quad n = 1, 2, \dots, N \end{aligned} \quad (17)$$

We give in Alg. 3 this particular version of the BC-VMFB algorithm which is called Regularized-GMCA

Algorithm 3: Regularized-GMCA

Initialization : $\alpha_1 \in \mathbb{C}^{N \times B}$, $\mathbf{A}_1 \in \mathbb{R}^{M \times N}$, $L_1 = \|\mathbf{A}_1\|^2$, $k = 1$;

repeat

$$\begin{aligned} \nabla_{\alpha} f(\mathbf{A}_k, \alpha_k) &= -\mathbf{A}_k^T (\mathbf{x} - \mathbf{A}_k \alpha \Phi^*) \Phi; \\ \alpha_{k+1} &= \text{prox}_{\frac{\psi}{L_k}}(\alpha_k - \frac{\nabla_{\alpha} f(\mathbf{A}_k, \alpha_k)}{L_k}); \\ \mathbf{A}_{k+1/2} &= \mathbf{x}\Phi\alpha_{k+1}^*; \\ \mathbf{a}_{k+1,n} &= \mathbf{a}_{k+1/2,n}/\|\mathbf{a}_{k+1/2,n}\|, \quad n = 1, 2, \dots, N; \\ L_{k+1} &= \|\mathbf{A}_{k+1}\|_F^2; \\ k &= k + 1; \end{aligned}$$

until convergence;

Next Proposition shows that this algorithm will enforce numerical decorrelation in over-determined noiseless setting

Proposition 1. *Suppose that $\mathbf{s} = \alpha\Phi^*$ where α is the minimizer of (14) by using the Algorithm 3 depending on λ and $M \geq N$, then $\mathbf{s}\mathbf{s}^T = \mathbf{D}$ as $\lambda \rightarrow 0$, where \mathbf{D} is a diagonal matrix*

The proof is postponed in Appendix VII.

The setting $\lambda \rightarrow 0$ corresponds to the noiseless scenario and the fact that this algorithm enforces decorrelation in the over-determined noiseless case will also be supported by the experiments in Section V.

Next, we present a more rigorous algorithm to combine sparsity and decorrelation.

B. A new functional involving sparsity and decorrelation

We reformulate the separation problem as a constrained optimization derived from (8) by turning the assumption into the constraint

$$\begin{cases} \arg \min_{\mathbf{A}, \alpha} \frac{1}{2} \|\mathbf{x} - \mathbf{A}\alpha\Phi^*\|_F^2 + \lambda \|\alpha\|_1 + \iota_C(\mathbf{A}) \\ \text{s.t. } \alpha\Phi^*\Phi\alpha^* = \mathbf{D} \end{cases} \quad (18)$$

We apply in this paper the Linearized Preconditioned Alternating Direction Method of Multipliers (LPADMM) [30] to solve problem (18). ADMM is a classic method to solve convex optimization problems. The method is based on the

alternating optimization of the corresponding augmented Lagrangian function. 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. In fact, the non-convex ADMM has already been used (for example [11]), however the convergence study of ADMM is limited in convex setting. In this article, despite the lack of convergence proof, the numerical experiments will illustrate its efficiency and correctness.

We first present the following proposition for Minimized Mean-Squared Error (MMSE) decorrelation transformation, which is a consequence of the result presented in [20] [16]. One can refer to Appendix VII for the proof.

Proposition 2. *Let $\mathbf{s} \in \mathbb{R}^{N \times T}$ be a 0-mean signal matrix with a positive definite covariance matrix $\Sigma_{\mathbf{s}} = \mathbf{s}\mathbf{s}^T$. Let \mathbf{W} be the optimal decorrelation transformation that minimizes the Mean-Squared Error (MSE) between the input \mathbf{s} and the output $\mathbf{y} = \mathbf{W}\mathbf{s}$ with covariance $\Sigma_{\mathbf{y}}$ being a diagonal matrix:*

$$\min_{\mathbf{y}=\mathbf{W}\mathbf{s}} \|\mathbf{s} - \mathbf{y}\|_F^2 \quad (19)$$

then $\mathbf{W} = \text{diag}(\Sigma_{\mathbf{s}}^{1/2}) \Sigma_{\mathbf{s}}^{-1/2}$ where $\text{diag}(\cdot)$ is the diagonal matrix formed with its diagonal elements.

We rewrite problem (18) as follow

$$\begin{cases} \arg \min_{\mathbf{A}, \alpha, \mathbf{s}} \frac{1}{2} \|\mathbf{x} - \mathbf{A}\alpha\Phi^*\|_F^2 + \lambda \|\alpha\|_1 + \iota_C(\mathbf{A}) + \iota_{\text{deco}}(\mathbf{s}) \\ \text{s.t. } \mathbf{s} = \alpha\Phi^* \end{cases} \quad (20)$$

where the function ι_{deco} is an indicator function that forces the decorrelation of \mathbf{s} that is:

$$\iota_{\text{deco}}(\mathbf{s}) = \begin{cases} 0 & \text{if } \mathbf{s}\mathbf{s}^T = \mathbf{D}, \text{ with } \mathbf{D} \text{ diagonal} \\ +\infty & \text{otherwise} \end{cases} \quad (21)$$

We formulate the augmented Lagrangian as follow :

$$\begin{aligned} L_k(\mathbf{A}, \alpha, \mathbf{s}, \eta) &= \frac{1}{2} \|\mathbf{x} - \mathbf{A}\alpha\Phi^*\|_F^2 + \lambda \|\alpha\|_1 + \iota_{\text{deco}}(\mathbf{s}) \\ &\quad + \langle \eta, \mathbf{s} - \alpha\Phi^* \rangle + \frac{\gamma}{2} \|\mathbf{s} - \alpha\Phi^*\|_F^2 \end{aligned} \quad (22)$$

where η is the dual variable. The LPADMM algorithm minimizes the augmented Lagrangian by iteratively updating the primal and dual variables:

$$\alpha_{k+1} = \arg \min_{\alpha} \langle \nabla_{\alpha} F_k(\alpha_k), \alpha \rangle + \lambda \|\alpha\|_1 + \frac{L_{F,k}}{2} \|\alpha_k - \alpha\|_F^2 \quad (23)$$

$$\mathbf{s}_{k+1} \in \arg \min_{\mathbf{s}} \langle \eta_k, \mathbf{s} - \alpha_{k+1}\Phi^* \rangle + \frac{\gamma}{2} \|\mathbf{s} - \alpha_{k+1}\Phi^*\|_F^2 + \iota_{\text{deco}}(\mathbf{s}) \quad (24)$$

$$\mathbf{A}_{k+1} = \arg \min_{\mathbf{A}} \frac{1}{2} \|\mathbf{x} - \mathbf{A}\alpha_{k+1}\Phi^*\|_F^2 + \iota_C(\mathbf{A}) \quad (25)$$

$$\eta_{k+1} = \eta_k + \gamma(\mathbf{s}_{k+1} - \alpha_{k+1}\Phi^*) \quad (26)$$

In (23),

$$F_k(\boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{x} - \mathbf{A}_k \boldsymbol{\alpha} \Phi^*\|_F^2 + \langle \eta_k, \mathbf{s}_k - \boldsymbol{\alpha} \Phi^* \rangle + \frac{\gamma}{2} \|\mathbf{s}_k - \boldsymbol{\alpha} \Phi^*\|_F^2$$

$L_{F,k}$ being the Lipschitz constant of $F_k(\boldsymbol{\alpha})$, and

$$\nabla_{\boldsymbol{\alpha}} F_k(\boldsymbol{\alpha}) = -\mathbf{A}_k^T (\mathbf{x} - \mathbf{A}_k \boldsymbol{\alpha} \Phi^*) \Phi - \eta_k \Phi - \gamma (\mathbf{s}_k - \boldsymbol{\alpha} \Phi^*) \Phi.$$

This sub-problem can be solved efficiently by Forward-backward splitting algorithm [15] and can be rewritten as follow:

$$\boldsymbol{\alpha}_{k+1} = \text{prox}_{\frac{\gamma}{L_k}} \left(\boldsymbol{\alpha}_k - \frac{\nabla_{\boldsymbol{\alpha}} F_k(\boldsymbol{\alpha}_k)}{\gamma} \right) \quad (27)$$

where prox is the proximal operator corresponding to ℓ_1 norm which is the Soft-thresholding operator.

The sub-problem (24) can be formulated as a decorrelation projection.

$$\begin{cases} \mathbf{s}_{k+1} = \underset{\mathbf{s}}{\text{argmin}} \frac{\gamma}{2} \left\| \mathbf{s} - \boldsymbol{\alpha}_{k+1} \Phi^* + \frac{\eta_k}{\gamma} \right\|_F^2 \\ \text{s.t. } \mathbf{s} \mathbf{s}^T = \mathbf{D} \end{cases} \quad (28)$$

and can be solved thanks to Proposition 2.

It is well known that ADMM type methods converge very slowly with fixed penalty parameter γ and several strategies of adaptive parameter have been proposed for different ADMM type methods. We follow [9] which balances, in each iteration, the primal residual and the dual residual. The complete algorithm is given in Algorithm 4.

Algorithm 4: LPADMM

Initialization : $\boldsymbol{\alpha}_1 \in \mathbb{C}^{N \times B}$, $\mathbf{s}_1 \in \mathbb{R}^{N \times T}$, $\mathbf{A}_1 \in \mathbb{R}^{M \times N}$, $L_1 = \|\mathbf{A}_1\|_F^2 + \gamma$, $k = 1$;

repeat

$$\begin{aligned} \nabla_{\boldsymbol{\alpha}} f(\mathbf{A}_k, \boldsymbol{\alpha}_k) &= -\mathbf{A}_k^T (\mathbf{x} - \mathbf{A}_k \boldsymbol{\alpha} \Phi^*) \Phi; \\ \boldsymbol{\alpha}_{k+1} &= \text{prox}_{\frac{\psi}{L_k}} \left(\boldsymbol{\alpha}_k - \frac{\nabla_{\boldsymbol{\alpha}} f(\mathbf{A}_k, \boldsymbol{\alpha}_k)}{L_k} \right); \\ \mathbf{s}_{k+1/2} &= \boldsymbol{\alpha}_{k+1} \Phi^* - \eta_{k+1} / \gamma; \\ \Sigma_{\mathbf{s}} &= \mathbf{s}_{k+1/2} \mathbf{s}_{k+1/2}^T; \\ \mathbf{W}_{\mathbf{s}} &= \text{diag}(\Sigma_{\mathbf{s}}^{1/2}) \Sigma_{\mathbf{s}}^{-1/2}; \\ \mathbf{s}_{k+1} &= \mathbf{W}_{\mathbf{s}} \mathbf{s}_{k+1/2}; \\ \mathbf{A}_{k+1/2} &= \mathbf{x} \Phi \boldsymbol{\alpha}_{k+1}^*; \\ \mathbf{a}_{k+1,n} &= \mathbf{a}_{k+1/2,n} / \|\mathbf{a}_{k+1/2,n}\|, \quad n = 1, 2, \dots, N; \\ \eta_{k+1} &= \eta_k + \gamma (\mathbf{s}_{k+1} - \boldsymbol{\alpha}_{k+1} \Phi^*) \\ L_{k+1} &= \|\mathbf{A}_{k+1}\|_F^2 + \gamma; \\ k &= k + 1; \end{aligned}$$

until convergence;

Although the linearized method and the adaptive parameter strategies can accelerate the convergence rate, the ADMM type methods still suffer from the slow convergence. Next, we propose a method that enforces decorrelation based on GMCA-like method by introducing a projection step which ensures the decorrelation of the sources during the procedure.

C. A decorrelated GMCA

Based on Regularized GMCA Alg 3, we present here an algorithm which enforces decorrelation in all scenarios by implementing the decorrelation projection at each iteration for

estimating the mixing matrix. The idea is that, in the step of minimizing with respect to the mixing matrix \mathbf{A} , before fixing the estimated sources \mathbf{s} , we perform a decorrelation projection to the estimated \mathbf{s} . By incorporating such a projection in the Regularized GMCA, we obtain the decorrelated-GMCA Alg. 5.

Algorithm 5: Decorrelated-GMCA

Initialization : $\boldsymbol{\alpha}_1 \in \mathbb{C}^{N \times B}$, $\mathbf{A}_1 \in \mathbb{R}^{M \times N}$, $L_1 = \|\mathbf{A}_1\|_F^2$, $k = 1$;

repeat

$$\begin{aligned} \nabla_{\boldsymbol{\alpha}} f(\mathbf{A}_k, \boldsymbol{\alpha}_k) &= -\mathbf{A}_k^T (\mathbf{x} - \mathbf{A}_k \boldsymbol{\alpha} \Phi^*) \Phi; \\ \boldsymbol{\alpha}_{k+1} &= \text{prox}_{\frac{\psi}{L_k}} \left(\boldsymbol{\alpha}_k - \frac{\nabla_{\boldsymbol{\alpha}} f(\mathbf{A}_k, \boldsymbol{\alpha}_k)}{L_k} \right); \\ \mathbf{s} &= \boldsymbol{\alpha}_{k+1} \Phi^*; \\ \mathbf{W}_{\mathbf{s}} &= \text{diag}(\Sigma_{\mathbf{s}}^{1/2}) \Sigma_{\mathbf{s}}^{-1/2}; \\ \mathbf{y} &= \mathbf{W}_{\mathbf{s}} \mathbf{s}; \\ \mathbf{A}_{k+1/2} &= \mathbf{x} \mathbf{y}^T; \\ \mathbf{a}_{k+1,n} &= \mathbf{a}_{k+1/2,n} / \|\mathbf{a}_{k+1/2,n}\|, \quad n = 1, 2, \dots, N; \\ L_{k+1} &= \|\mathbf{A}_{k+1}\|_F^2; \\ k &= k + 1; \end{aligned}$$

until convergence;

output: $\mathbf{A}_k \mathbf{s} = \mathbf{W}_{\mathbf{s}} \mathbf{s}$

It is worthy mentioning that these extra steps will not bring too much computational cost because of the small size of $\Sigma_{\mathbf{s}}$. Although, the convergence proof of this algorithm remains an open problem especially because of the decorrelation projection, the experiments in section V will illustrate its good performance. We must stress that the decorrelation step is useful only for the estimation of the mixing matrix. Indeed, applying such a projection also for the time-frequency coefficient leads to lower performances in practice.

D. Tuning the parameters

We have proposed three algorithms that combine decorrelation and sparsity in different ways. However, the practical implement of these algorithms encounters the parameter choosing problem. As the hyper-parameter λ controls the tradeoff between the fidelity term and the penalty term, the choice of this hyper-parameter plays a crucial role in the separation process. As mentioned before, in the noiseless case, setting $\lambda \rightarrow 0$ is a sound choice. However, for small λ , the algorithms require a large number of iterations to reach convergence. Therefore, in practice, the continuation trick also known as warm-start or fixed point continuation [22] is used: we first run the algorithm with a large value of λ which is then iteratively decreased while the algorithm is initialized with the result of the previous run.

V. EXPERIMENTS

After the presentation of the experimental setup, we describe in this section the performances of proposed methods in different scenarios.

A. Experimental setup

For all the experiments, we have taken ten signals from the SISEC 2011 [2] data base. The STFT was computed with half-overlapping tight Hann window using the ltfat toolbox [33].

The separation performances were assessed using the Signal to Distorsion Ratio (SDR) and Signal to Interference Ratio (SIR) [34]. 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. All the experiments are initialized by random signals. In each experiment, we apply the warm-start technique and choose the hyperparameter λ which corresponds to the best SDR.

To better evaluate the performance of the proposed algorithms, we present also the performance of the oracle setting which means that the mixing matrix is supposed to be known. We denote the oracle setting of the algorithm based only on sparsity as *Oracle*, and the one based on both sparsity and decorrelation by using LPADMM method, as *LPADMM Oracle*. In order to stress that the decorrelation should be taken into account during the procedure, a whitened post-processing is implemented in every experiment for *PALM*, *Regu-GMCA* and *Oracle* algorithms.

B. Over-determined BSS

We present here the performances in the determined ($M = 3$, $N = 3$) and over-determined ($M = 5$, $N = 3$) cases with different input Signal to Noise Ratio (SNR). The mixing matrices are created artificially with condition numbers equal to 14.8 and 6.1 respectively. EFICA [13] and CoM2 [16] are used as a baseline for ICA. The performances are shown in Fig. 1 and Fig. 2.

In both cases, although the ICA methods (*EFICA*, *CoM2*) outperform other algorithms when the input noise is negligible, their performances collapse rapidly as the noise increases, and become outperformed by other approaches. We also notice that the SDR of ICA methods decreases more rapidly than SIR, which indicates that ICA methods have good separation performance but behave poorly when dealing with noise.

Moreover, one can remark that *LPADMM*, *Deco-GMCA* and *Regu-GMCA* perform similar in both SDR and SIR and outperform GMCA-like method and the *PALM* minimization of Alg. 1, especially when the input SNR is large. This observation confirms that the combination of decorrelation and sparsity can improve the separation performance. It is also important to mention that *PALM* algorithm performs better than *GMCA* in terms of SDR in the determined case, while it performs better in both SDR and SIR in the over-determined case. At last, if we do the comparison between the determined and the over-determined case, we can see the improvement from determined to over-determined, which is a natural effect due to the extra information given in the over-determined case.

Compared with the "oracle" setting, we can see that although *Oracle* and *LPADMM oracle* have better performances, differences are very small. It is worthy mentioning that these differences reduce to almost zero when the noise is of large energy. This observation shows that the redundancy given in the (over-)determined case decreases the importance of the information given by the mixing matrix.

C. Under-determined BSS

1) *Performance analysis as a function of input SNR*: The proposed algorithms are evaluated in the under-determined case ($M = 2$, $N = 3$) with different input SNR using the mixing matrix from [2] whose condition number is 4.2. In this case, *DUET* algorithm is used as a baseline. The SDR/SIR performances are displayed on Fig. 3. The "oracle" setting is displayed on the same figure. As *GMCA* fails the separation in this case, we do not show its performance.

From Fig. 3 we see that, *LPADMM*, *Regu-GMCA* and *Deco-GMCA* have similar performances and outperform *PALM* in terms of SDR. Moreover, as expected, *LPADMM* outperforms other approaches in SIR when the noise is of small energy. It is interesting to notice that *Regu-GMCA* has relatively poor performances in SIR compared to *LPADMM* and *Deco-GMCA* while it is outperformed by *PALM* when the input SNR is large. At last, it is important to notice that there is a relatively large gap between the performances of the proposed methods and the State-of-art method *DUET*.

By considering the "oracle" setting, one can see that *LPADMM Oracle* outperforms *Oracle* and the difference is relatively large in terms of SIR when the input noise is small. This observation shows that the extra decorrelation indeed improves separation performances. Moreover, there is an obvious gap between the proposed methods and the "oracle" settings. This remark illustrates that the information given by the mixing matrix becomes more important in the under-determined case.

Furthermore, one can notice that for *Oracle*, *Regu-GMCA* and *Deco-GMCA*, the SIR decreases as the input SNR increases. A possible explanation is the choice of the hyperparameter λ : for the separation with relatively small input SNR, we often choose a relatively larger λ which will lead to "sparser" solution and will naturally have relatively higher SIR.

2) *Performance analysis as a function of the condition number of the mixing matrix*: Now we show the performances of the proposed algorithms as a function of the condition number of the mixing matrix on Fig. 4. A large condition number indicates a poor mixing situation (either the microphones or the sources are placed close to one another). For the sake of clarity, we did not add any noise.

One can see on Fig. 4 that, except *DUET*, the performances of all the other algorithms decrease in terms of SDR as the condition number of the mixing matrix augments. Secondly, *LPADMM*, *Regu-GMCA* and *Deco-GMCA* outperform *PALM* algorithm in both SDR and SIR for all mixing condition. It is also interesting to notice that *LPADMM* is relatively robust to the condition number of the mixing matrix in SIR. At last, it can be seen from the figure that the Time-frequency masking method *DUET* is not sensible to the condition number of the mixing matrix.

3) *Performance analysis as a function of the number of sources*: We show in Fig. 5 the performances of the proposed algorithms as a function of the number of sources. In all cases, the number of observations is set to $M = 2$. We did not add any noise in these experiments.

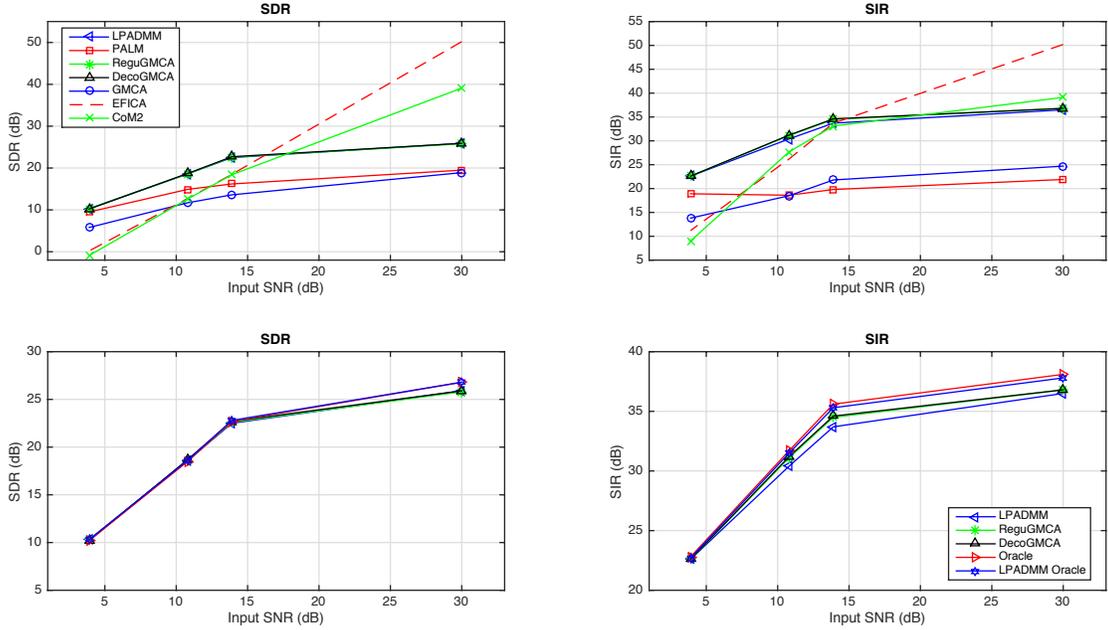


Fig. 1. Top: Performances evaluation of all algorithms in the determined case with $M = 3$ and $N = 3$. Bottom: Proposed algorithms vs oracle setting

From the figure, first, one can notice that the performances of all methods decrease rapidly as the number of sources increases. It is natural because as the number of sources grows, the separation problem becomes more and more ill-posed. Secondly, one can remark that when the number of sources is large, *LPADMM* outperforms *Regu-GMCA* and they both perform better than *Deco-GMCA* which decreases relatively rapidly in terms of SIR. At last, it can be seen that the gap between the DUET method and the proposed methods increases as the number of sources grows. We can see that the two "oracle algorithms" perform very similarly.

D. Computational comparison

We end this experimental section by giving some indications about the computational time of different methods. Fig. 6 shows the separation performance in terms of the average computational time in the under-determined noiseless case. It can be noticed that the performance of *Deco-GMCA* increases rapidly in term of SDR, but the SIR start decreasing while the SDR reach its maximum. This suggests that a better choice of the hyperparameter λ could be done instead of $\lambda \rightarrow 0$. It is also obvious that *LPADMM* converge very slowly and is relatively not stable.

VI. DISCUSSION AND CONCLUSION

We studied and proposed several iterative methods to separate instantaneous mixture in under-determined and over-determined BSS. In particular, we developed three approaches to combine both the decorrelation in the time domain and the sparsity in the time-frequency domain of the sources. It has been shown that these approaches lead to better separation

performances in both SDR and SIR, and is relative robust to noise.

In all the three methods, we chose to maintain the energy of the sources, and forced each columns of the mixing matrix to have a unit norm, in order to deal with the scaling ambiguity. In theory, it is totally possible to force the sources to have a unit norm, and then to have $\mathbf{ss}^T = \mathbf{I}$ to release the constraint on the norm of the column of the mixing matrix. Surprisingly enough, this approach does not work as well as the proposed approach.

Future works will focus on extending the decorrelation and sparsity combination to convolutive mixture. Structured sparsity will also be considered to take signal structures into account. From an applicative point of view, such an approach may also be applied to M/EEG and fMRI signals.

VII. APPENDIX

We first give the proof for Proposition 1

Proof. First we reformulate (14) as a constrained optimization problem as $\lambda \rightarrow 0$:

$$\begin{cases} \min_{\alpha, \mathbf{A}} \|\alpha\|_1 + \iota_C(\mathbf{A}) \\ \text{s.t. } \mathbf{x} = \mathbf{A}\alpha\Phi^* \end{cases} \quad (29)$$

Then if α and \mathbf{A} is a minimiser of (29), one must have that $\mathbf{x} = \mathbf{A}\alpha\Phi^*$. On the other hand, at convergence, step 2 and 3 of Algo. 3 ensures that $\mathbf{A} = \mathbf{x}\Phi\alpha^*\mathbf{D}$ where \mathbf{D} is a diagonal matrix. Then one can get $\mathbf{ss}^T = \alpha\Phi^*\Phi\alpha^* = \mathbf{D}^{-1}$ which is a diagonal matrix if $M \geq N$ by simple linear algebra. \square

Now we give the proof for Proposition 2

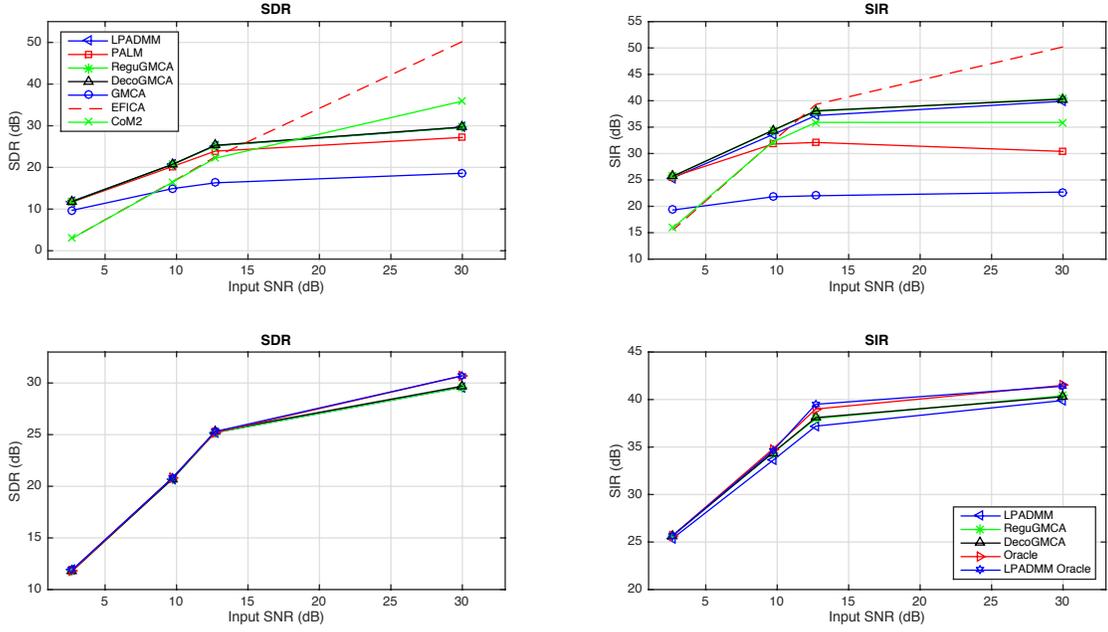


Fig. 2. Top: Performances evaluation of all algorithms in determined case with $M = 5$ and $N = 3$. Bottom: Proposed algorithms vs oracle setting

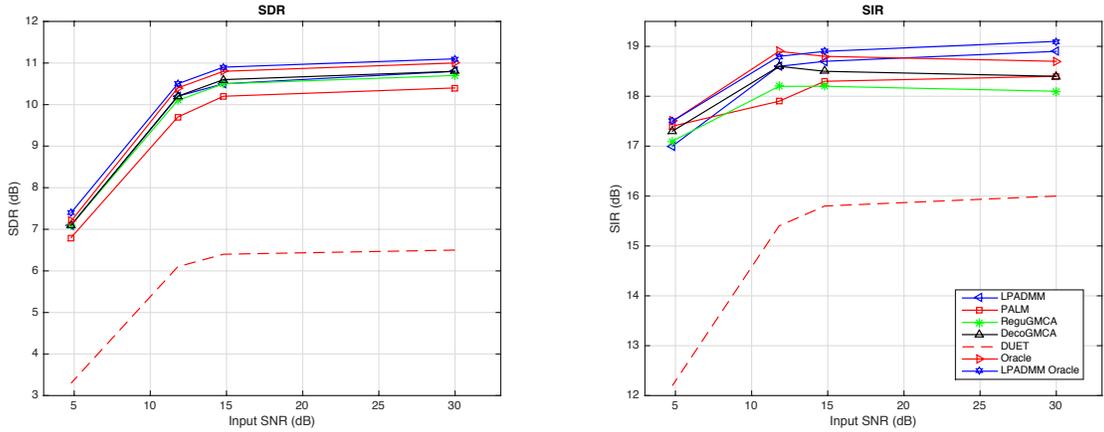


Fig. 3. Performances evaluation in under-determined case with $M = 2$ and $N = 3$

Proof. We consider the problem (19) in the following way:

$$\begin{aligned} \min_{\bar{\mathbf{W}}} \|\mathbf{s} - \bar{\mathbf{W}}\bar{\mathbf{y}}\|_F^2 \\ \text{s.t. } \bar{\mathbf{W}}\bar{\mathbf{W}}^T = \mathbf{D} \end{aligned} \quad (30)$$

where $\bar{\mathbf{y}}$ is the whitened signal, i.e. $\bar{\mathbf{y}} = \Sigma_s^{-1/2}\mathbf{s}$ and \mathbf{D} is any diagonal matrix. Then the MSE can be rewritten as :

$$e = \|\mathbf{s}\|_F^2 + \|\bar{\mathbf{W}}\bar{\mathbf{y}}\|_F^2 - 2\langle \mathbf{s}, \bar{\mathbf{W}}\bar{\mathbf{y}} \rangle \quad (31)$$

According to Cauchy-Schwarz inequality, one has $\langle \mathbf{s}, \bar{\mathbf{W}}\bar{\mathbf{y}} \rangle \leq \|\mathbf{s}\|_F \|\bar{\mathbf{W}}\bar{\mathbf{y}}\|_F$ with equality if and only if $\bar{\mathbf{W}}$ is a diagonal matrix. We note \bar{d}_i the i -th element on the diagonal of $\bar{\mathbf{W}}$, then we can reformulate e as :

$$e = \sum_{i=1}^N (\mathbf{s}_i - \bar{d}_i \bar{\mathbf{y}}_i)(\mathbf{s}_i - \bar{d}_i \bar{\mathbf{y}}_i)^T \quad (32)$$

where \mathbf{s}_i and $\bar{\mathbf{y}}_i$ is the i -th row of \mathbf{s} and $\bar{\mathbf{y}}$ respectively. Now minimizing e with respect to \bar{d}_i is a simple problem with the solution $\bar{d}_i = \mathbf{s}_i \bar{\mathbf{y}}_i^T$ which means $\bar{\mathbf{W}} = \text{diag}(\mathbf{s}\bar{\mathbf{y}}^T)$. \square

One can notice that one of the conditions of this proposition is that Σ_s is a positive definite matrix. However, it is not always the case in practice. The strategy we use is that we simply do not do any decorrelation projection if Σ_s is not positive definite. One explanation is that we assume that Σ_s will always be positive definite after a finite number of iterations.

REFERENCES

- [1] Bernard Ans, Jeanny Héroult, and Christian Jutten. Adaptive neural architectures: detection of primitives. *Proc. of COGNITIVA85*, pages 593–597, 1985.

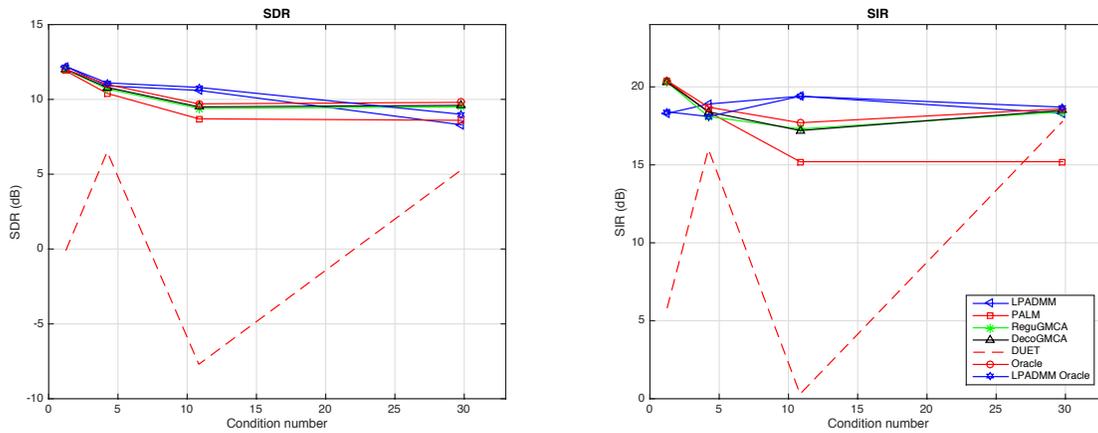


Fig. 4. SDR/SIR vs condition number of the mixing matrix

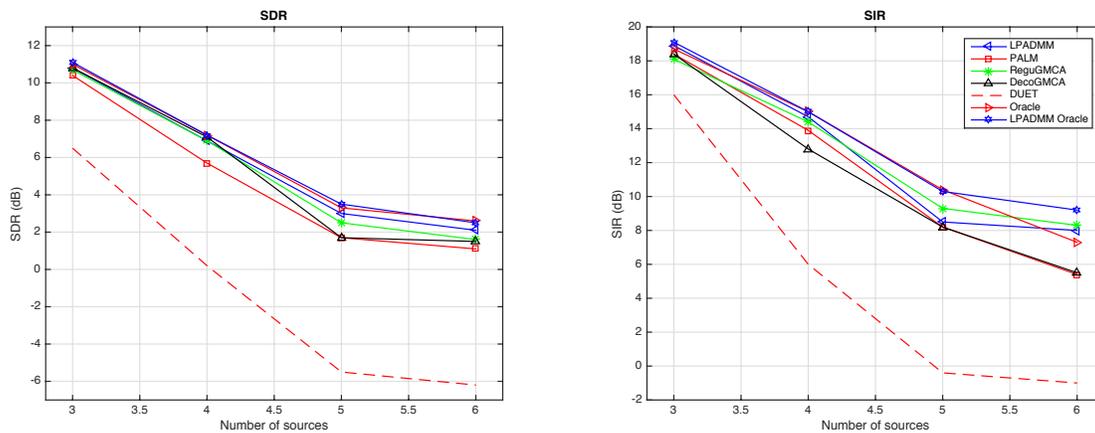


Fig. 5. SDR/SIR vs condition number of the number of source

- [2] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe, and A. Benichoux. The 2011 signal separation evaluation campaign (SiSEC2011): audio source separation. In *Latent Variable Analysis and Signal Separation*, pages 414–422. Springer, 2012.
- [3] P. Balazs, M. Dörfler, M. Kowalski, and B. Torrèsani. Adapted and adaptive linear time-frequency representations: a synthesis point of view. *Signal Processing Magazine, IEEE*, 30(6):20–31, 2013.
- [4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [5] J. Bobin, J.-L. Starck, J. Fadili, and Y. Moudden. Sparsity and morphological diversity in blind source separation. *Image Processing, IEEE Transactions on*, 16(11):2662–2674, 2007.
- [6] Jérôme Bobin, Jean-Luc Starck, Yassir Moudden, and Mohamed Jalal Fadili. Blind source separation: the sparsity revolution. *Advances in Imaging and Electron Physics*, 152:221–302, 2008.
- [7] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, pages 1–36, 2013.
- [8] H. Boumaraf. *Séparation aveugle de mélanges convolutifs de sources*. PhD thesis, Université Joseph-Fourier-Grenoble I, 2005.
- [9] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
- [10] Jean-Francois Cardoso. Infomax and maximum likelihood for blind source separation. 1997.
- [11] Rick Chartrand and Brendt Wohlberg. A nonconvex admm algorithm for group sparsity with sparse groups. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6009–6013. IEEE, 2013.
- [12] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. A block coordinate variable metric forward-backward algorithm. <https://hal.archives-ouvertes.fr/hal-00945918/>, 2013.
- [13] A. Cichocki, S.-I. Amari, K. Siwek, T. Tanaka, A. H. Phan, R. Zdunek, S. Cruces, P. Georgiev, Y. Washizawa, and Z. Leonowicz. Icalab toolboxes. URL: <http://www.bsp.brain.riken.jp/ICALAB>, 2007.
- [14] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.
- [15] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, November 2005.
- [16] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.
- [17] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.
- [18] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D’ardenne, W. Richter, J.D. Cohen, and J. Haxby. Independent component analysis for brain fmri does not select for independence. *Proceedings of the National Academy of Sciences*, 106(26):10415–10422, 2009.
- [19] Mike Davies. Identifiability issues in noisy ica. *IEEE Signal processing letters*, 11(5):470–473, 2004.
- [20] Y. C Eldar and A. V. Oppenheim. MMSE whitening and subspace whitening. *Information Theory, IEEE Transactions on*, 49(7):1846–1851, 2003.

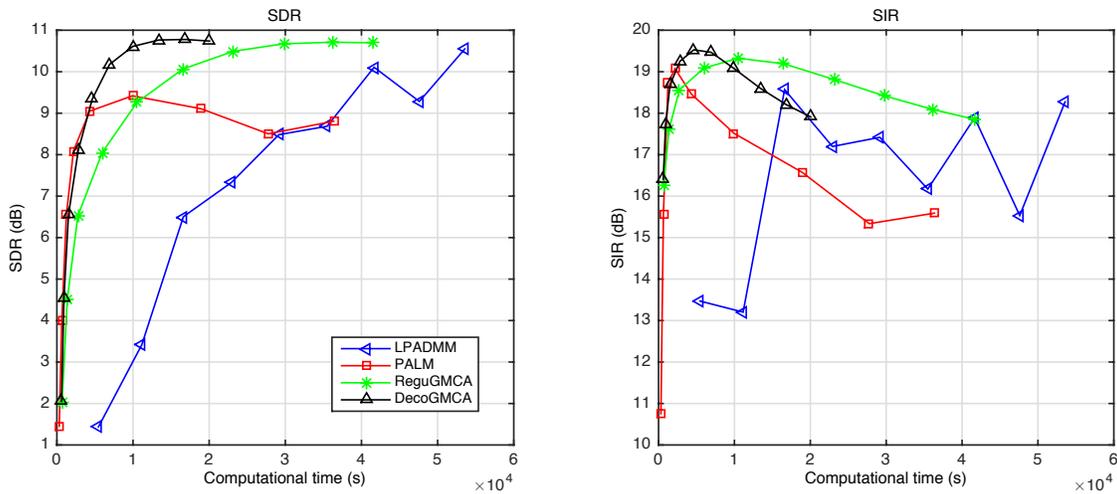


Fig. 6. Computational time evaluation in under-determined noiseless case

- [21] Fangchen Feng and Matthieu Kowalski. An unified approach for blind source separation using sparsity and decorrelation. In *Eusipco 2015*, 2015.
- [22] Elaine T Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for ℓ_1 -minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.
- [23] J Hérault. B. ans, réseaux de neurones à synapses modifiables: Decodage de messages sensoriels composites par une apprentissage non supervisé et permanent. *CR Acad. Sci. Paris, vol. sCrie*, 111:277–282.
- [24] Jeanny Hérault, Christian Jutten, and Bernard Ans. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *10 Colloque sur le traitement du signal et des images, FRA, 1985*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 1985.
- [25] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.
- [26] Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.
- [27] Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin J McKeown, Vicente Iragui, and Terrence J Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(02):163–178, 2000.
- [28] Te-Won Lee, Mark Girolami, Anthony J Bell, and Terrence J Sejnowski. A unifying information-theoretic framework for independent component analysis. *Computers & Mathematics with Applications*, 39(11):1–21, 2000.
- [29] Martin J McKeown, Terrence J Sejnowski, et al. Independent component analysis of fmri data: examining the assumptions. *Human brain mapping*, 6(5-6):368–372, 1998.
- [30] Yuyuan Ouyang, Yunmei Chen, Guanghui Lan, and Eduardo Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. *arXiv preprint arXiv:1401.6607*, 2014.
- [31] MD Plumbley, SA Abdallah, JP Bello, ME Davies, J Klingseisen, G Monti, and MB Sandler. Ica and related models applied to audio analysis and separation. In *In Proc. 4th Int. ICSC Symposium on Soft Computing and Intelligent Systems for Industry*. Citeseer, 2001.
- [32] H. Sawada, S. Araki, and S. Makino. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):516–527, 2011.
- [33] P. L. Sondergaard, B. Torrèسانی, and P. Balazs. The linear time frequency analysis toolbox. *International Journal of Wavelets, Multiresolution and Information Processing*, 10(04), 2012.
- [34] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469, 2006.
- [35] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *Signal Processing, IEEE transactions on*, 52(7):1830–1847, 2004.
- [36] Michael Zibulevsky and Michael Elad. L1-l2 optimization in signal and image processing. *Signal Processing Magazine, IEEE*, 27(3):76–88, 2010.
- [37] Michael Zibulevsky and Barak A Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882, 2001.