Unsupervised separation of sparse sources in the presence of outliers

Cécile Chenot and Jérôme Bobin

***Abstract*—Unsupervised or Blind Source Separation (BSS) aims at recovering underlying sources from linear mixtures. In real world applications, these multichannel measurements can contain some corrupted entries which are detrimental for most of the BSS techniques. We first discuss how the presence of outliers alters the sparse BSS problem and point out that the outliers can be detected and extracted thanks to sparsity. Building upon the rationale of the AMCA algorithm, which has been developed to process partially correlated sources, we propose to penalize and detect the outliers simultaneously with the estimation of the sources and the mixing matrix. Numerical experiments illustrate the robustness of this new approach on a wide range of applications, including the determined case.**

***Index Terms*—Blind source separation, sparse representations, sparsity, robust recovery, outliers.**

I. Introduction

BLIND Source Separation (BSS) is a separation components procedure well-suited to extract meaningful information from multichannel data. The interest for this multivariate data processing method is motivated by its numerous applications in various fields such as astrophysics [1] or hyperspectral unmixing [2] to only name a few.

In the noiseless BSS context, the m multichannel observations $\{\mathbf{X}_i\}_{i=1..m}$ are assumed to be the linear mixture of $n \leq m$ sources $\{\mathbf{S}_j\}_{j=1..n}$ with $t > m$ samples. This model can be conveniently recast with the following matrix form:

$$\mathbf{X} = \mathbf{AS},$$

where $\mathbf{X} \in \mathbf{R}^{m\times t}$ stands for the observations, $\mathbf{A} \in \mathbf{R}^{m\times n}$ the mixing matrix, and $\mathbf{S} \in \mathbf{R}^{n\times t}$ the sources.

The aim of the BSS techniques is to estimate $\mathbf{A}$ and $\mathbf{S}$ from $\mathbf{X} = \mathbf{AS}$. This is a challenging task as the number of solutions $(\mathbf{A}, \mathbf{S})$ is infinite. Prior information on the sources or the mixing matrix need to be exploited to recover the original sources. Based on the prior used to unmix the sources, BSS techniques can be divided into three main classes.

In the well-known Independent Component Analysis (ICA) framework, the sources are assumed to be statistically independent and non-Gaussian [3]. The ICA techniques differ from each other on the way the independence of the sources is promoted: minimization of the mutual information using the Kullback-Leibler divergence for example [4]. In the Nonnegative Matrix Factorization (NMF) domain, both the sources and the mixing matrix are assumed to be nonnegative [5]. This assumption is naturally motivated by the physical aspect of some problems e.g. spectra unmixing or audio source separation. The last approach relies on the sparse modeling of the sources [6]–[8]. We will focus on this last approach in the remaining of the paper.

C.C and J. B with CEA, IRFU, Service d'Astrophysique - SEDI, 91191 Gif-sur-Yvette Cedex, France.

While real-world data are generally contaminated with noise, most BSS methods are sensitive to data corruption [9]. Model deviations from the above instantaneous linear model are usually represented by an additional Gaussian term $\mathbf{N}$. However, this term can only stand for small and dense noise whereas most of the applications are also frequently corrupted by rare and large errors. These spurious *outliers*, which will be designated by $\mathbf{O}$, include observed unexpected physical events or malfunctions of captors which can be represented by stripping noise or impulsive noise (see for instance [10]). As they correspond to infrequent errors, they can be considered to be sparse in the domain of observation. In order to account for both the presence of Gaussian noise and outliers, we will further consider that the observations can be expressed as:

$$\mathbf{X} = \mathbf{AS} + \mathbf{O} + \mathbf{N},$$

where $\mathbf{O} \in \mathbf{R}^{m\times t}$ stands for the outliers, and $\mathbf{N} \in \mathbf{R}^{m\times t}$ the Gaussian noise.

Robust BSS methods in the literature:

In the current litterature, only a few BSS methods have been developed to manage the presence of outliers. These techniques can be divided into three groups: the "two-steps "separation methods, the robust ICA techniques, and last, the component separation methods.

The two-steps methods: They consist in: i) eliminating the outliers from the observations in a first step, ii) and then performing the separation on the "outliers-free"data. The first denoising step is crucial as all the observations should be cleaned with a high precision in order to reliably perform the separation in a second step. Unfortunately, discriminating between $\mathbf{AS}$ and $\mathbf{O}$ is challenging to perform unless $m \gg n$. In this specific case, the authors of [11] have shown that it is indeed possible to separate the low-rank matrix $\mathbf{AS}$ from the sparse matrix $\mathbf{O}$ with a high accuracy, if the latter are randomly drawn. The so-called robust PCA method performs by minimizing:

$$\underset{\mathbf{R},\mathbf{O}}{\text{minimize}} \|\mathbf{R}\|_* + \lambda \|\mathbf{O}\|_1 \text{ subject to } \mathbf{X} = \mathbf{R} + \mathbf{O}, \quad (1)$$

where $\mathbf{R} = \mathbf{AS}$. This separation based on the difference of structure between a low-rank term $\mathbf{R}$ and a sparse and broadly distributed term has inspired several works, especially in hyperspectral image denoising where the assumption $m \gg n$ is valid [10], [12]. However, in addition to the optimal choice

of λ that is generally not trivial in practice, the low-rank assumption is not valid in a wide range of applications, especially when only few observations are available.

The robust ICA techniques: In the outliers-free setting, based on the assumptions that the sources are independent and that $m = n$, the ICA based methods look for an unmixing matrix $\mathbf{B}$ such that the estimated sources $\tilde{\mathbf{S}} = \mathbf{BX}$ are statistically independent [3]. One way to measure the independence of the estimated sources is to use the mutual information of the sources, defined as the Kullback-Leibler (KL) divergence between the product of their marginal distributions $\prod_{i=1}^{n} p_{\mathbf{S}}(\tilde{\mathbf{S}}_i)$ and their joint distribution $p_{\mathbf{S}}(\tilde{\mathbf{S}})$ [4]:

$$\mathbb{D}_{KL}(p_{\mathbf{S}}(\tilde{\mathbf{S}}) \| \prod_{i=1}^{n} p_{\mathbf{S}}(\tilde{\mathbf{S}}_i)) = \int p_{\mathbf{S}}(\tilde{\mathbf{S}}) \log \left(\frac{p_{\mathbf{S}}(\tilde{\mathbf{S}})}{\prod_{i=1}^{n} p_{\mathbf{S}}(\tilde{\mathbf{S}}_i)} \right) d\tilde{\mathbf{S}}.$$

The value of the KL-divergence is non-negative and equal to zero if and only the estimated sources are independent. Unfortunately, it is not robust to the presence of outliers [9]. To overcome this problem, the authors of [13] propose to replace the KL-divergence by the β-divergence, for which the influence of the outliers is limited [13]. Similarly, the β-divergence between the product of the marginal densities of the estimated sources and their joint density is non-negative and equal to zero if and only if the estimated sources are independent (for $\beta > 0$, otherwise, we refer to the KL divergence):

$$\mathbb{D}_{\beta}(p_{\mathbf{S}}(\tilde{\mathbf{S}}) \| \prod_{i=1}^{n} p_{\mathbf{S}}(\tilde{\mathbf{S}}_i)) = \frac{1}{\beta} \int p_{\mathbf{S}}(\tilde{\mathbf{S}}) \left(p_{\mathbf{S}}^{\beta}(\tilde{\mathbf{S}}) - \prod_{i=1}^{n} p_{\mathbf{S}}^{\beta}(\tilde{\mathbf{S}}_i) \right) d\tilde{\mathbf{S}} - \frac{1}{\beta+1} \int \left(p_{\mathbf{S}}^{\beta+1}(\tilde{\mathbf{S}}) - \prod_{i=1}^{n} p_{\mathbf{S}}^{\beta+1}(\tilde{\mathbf{S}}_i) \right) d\tilde{\mathbf{S}}.$$

The strategy proposed in [13] amounts to maximizing the quasi log-likelihood of the β-divergence between the empirical marginal densities of the estimated sources and their empirical joint density. It does not need any whitening of the data, but only that $m = n$. Nevertheless, the efficiency of this approach highly depends on the choice of the parameter β, which is generally problematic in practice.

The component separation methods: Last, the third strategy amounts to modeling the outliers as an extra component of the mixture model. This therefore requires estimating jointly the mixing matrix, the sources and the outliers. This method has been used in the NMF framework. The approaches in [14] and [15] are quite similar. They minimize a function of the form:

$$\underset{\mathbf{A}\geq 0, \mathbf{S}\geq 0, \mathbf{O}}{\text{minimize}} \frac{1}{2} \left\| \mathbf{X} - \mathbf{AS} - \mathbf{O} \right\|_F^2 + \lambda \left\| \mathbf{O} \right\|_p, \qquad (2)$$

where the first term is the data fidelity term and the second term enforces the sparsity of $\mathbf{O}$, with $p = 1$ in [14] and $p = (1, 2)$ in [15] (in these works, the outliers are assumed to be uniformly sparse). In [16], the authors propose to use a β-divergence instead of the Frobenius norm to overcome the fact that the residue $\mathbf{X} - \mathbf{AS} - \mathbf{O}$ may not be correctly handled with the Frobenius norm in some situations. Moreover, they assume that the outliers are also nonnegative and corrupt entirely some columns of $\mathbf{X}$. Consequently, they use the $\ell_{2,1}$ to promote the sparsity of the columns of $\mathbf{O}$. In a Bayesian framework, and based on a MCMC method, the authors of [17] proposed to estimate the three variables by taking also into account a possible correlation of the outliers, which are not assumed to be non-negative unlike [16].

We point out that without further assumption, the decomposition $\mathbf{X} = \mathbf{AS}$ where $\mathbf{A}$ and $\mathbf{S}$ are non-negative is not necessarily unique [18]. Some conditions, such as the presence of pure pixels in the observations, ensure the uniqueness of the solution and so the recoverability of the sources. However in some applications [1], such assumptions cannot be made, and it is then necessary to use another prior such as the independence of the sources or their sparsity in another domain to discriminate between the sources.

Contribution:

This article details the impact of the presence of outliers in sparse BSS problems and presents a new robust method able to manage the presence of sparsely corrupted data in a wide range of applications.

More precisely, it extends the work done in [19], in which a framework of sparse BSS in the presence of outliers and a robust BSS method, rGMCA, were introduced. Building upon sparsity and the morphological diversity principle, the algorithm rGMCA, whose basics are summarized in section III, has been shown to be able to estimate reliably the mixing matrix in the over-determined setting. However, it is much less successful in the determined setting or in the presence of a large amount of small outliers. In section IV, we introduce a new algorithm that is able to handle the presence of outliers in a large range of applications, including the determined case. This approach exploits the structure of the problem and its similarities with the problem of sparse BSS in the presence of partially correlated sources, for which the algorithm AMCA has been recently proposed in [20]. This algorithm, coined robust AMCA (rAMCA), is based on the parsimony of the sources and outliers but makes no assumption on the low-rankness or non-negativity of the mixture parameters. Besides, it does not requires any fine parameter tuning, and is thus simple to use. Numerical experiments, presented in the section V, illustrate the good performances of rAMCA for a broad variety of settings.

Notations

Uppercase boldface letters denote matrices. The Moore-Penrose inverse of the matrix $\mathbf{M}$ is designated by $\mathbf{M}^{\dagger}$. The jth column of $\mathbf{M}$ is denoted $\mathbf{M}^j$, the ith row $\mathbf{M}_i$, and the i, jth entry $\mathbf{M}_{i,j}$. The norm $\|\mathbf{M}\|_2$ denotes the Frobenius norm of $\mathbf{M}$, and more generally $\|\mathbf{M}\|_p$ designates the p-norm of the matrix $\mathbf{M}$ seen as a long vector. The soft-thresholding operator is denoted $\mathcal{S}_{\lambda}(\mathbf{M})$, where

$$[\mathcal{S}_{\lambda}(\mathbf{M})]_{i,j} = \begin{cases} \mathbf{M}_{i,j} - \text{sign}(\mathbf{M}_{i,j}) * \lambda_i \text{ if } |\mathbf{M}_{i,j}| > \lambda_i \\ 0 \text{ otherwise} \end{cases}$$

The operator mad designates the median absolute deviation, and last $\Pr$ stands for probability.

II. Sparse BSS

In this section, we start by reviewing the basics of sparse BSS. Then, we will discuss how the outliers impact sparse BSS.

A. Sparsity and BSS: the Morphological Diversity Principle

In the context of sparse BSS, the sources $\{\mathbf{S}\}_{i=1..n}$ are assumed to be sparse in a given dictionary $\mathbf{\Phi}$ or approximatively sparse in $\mathbf{\Phi}$:

$$\mathbf{S}_i = \alpha_i \mathbf{\Phi}, \, \forall i = 1..n,$$

where α_i is composed of few non zero coefficients, or respectively α_i is composed of negligible entries with few significant entries.

In the remaining of this paper, we will assume that the outliers and the sources have a similar morphology in the sense that they are sparse in a same dictionary $\mathbf{\Phi}$. For simplicity and without loss of generality, we will consider that the sources and the outliers are sparse in the direct domain ($\mathbf{\Phi} = \mathbf{I}$).

Sparsity has been shown to be an effective separation criterion in BSS [6], as it enhances the contrast between the sources. Indeed, in the transformed domain, the sources are represented with only few significant coefficients, which concentrate the information of each source. Then, the largest coefficients of each source, which are the most representative, are very likely to be located at different positions as the sources are distinct. These large coefficients, which are not active simultaneously in several sources are thus the most discriminant for the source separation. This fact has been designated by Morphological Diversity Principle (MDP) in [7].

The MDP is particularly useful for source separation. Indeed if some sources respecting the MDP are mixed, then the significant samples of $\mathbf{X}$ are placed along the directions given by the columns of $\mathbf{A}$. Furthermore, these significant entries which correspond to the contribution of the largest entries of $\mathbf{S}$ are shared by several measures and are thus non sparse. Thus, while the sources are not correctly unmixed, their corresponding mixtures are less sparse than the initial sources: to unmix sources respecting the MDP, one can look for the sparsest joint sources [7].

Building upon sparsity and the MDP, the algorithm GMCA [7] has been developed to estimate the tuple $(\mathbf{A}, \mathbf{S})$ from $\mathbf{X}$ by estimating the sparsest sources, measured by the ℓ_1-norm of the sources, by minimizing:

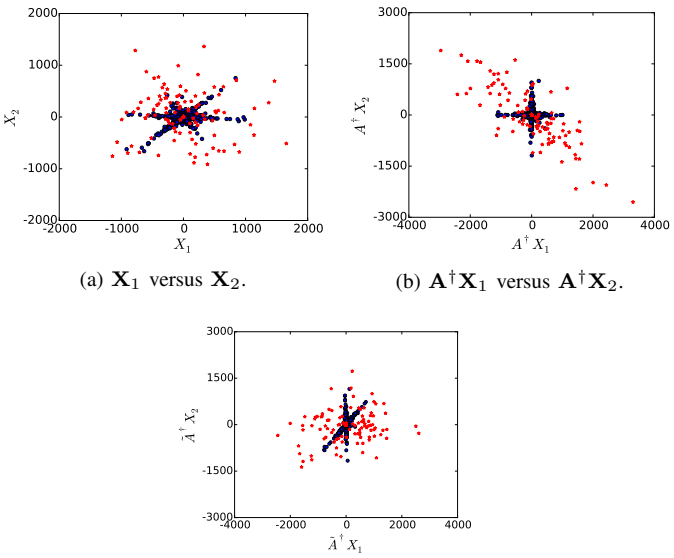
$$\underset{\mathbf{S},\mathbf{A}}{\text{minimize}} \, \frac{1}{2} \left\| \mathbf{X} - \mathbf{A}\mathbf{S} \right\|_2^2 + \sum_{i=1}^{n} \lambda_i \left\| \mathbf{S}_i \right\|_1 ,$$

where the first term is the data-fidelity term and the second term promotes the sparsity of the sources.

B. Sparse BSS in the presence of outliers

The algorithm GMCA has been shown to be robust to a Gaussian noise [7] but it is however highly sensitive to the outliers [19]. In this paragraph, we will analyze more precisely how the presence of outliers impacts sparse BSS.

Throughout the remaining of this article, we will assume that the outliers are distributed in general position: they do not cluster in any specific direction. This implies that the outliers contribution in the source domain (by looking at $\mathbf{A}^\dagger\mathbf{O}$) will be also in general position. In the following, we assume that the outliers corrupt entirely some columns of $\mathbf{X}$ fig.1a so that the support of the columns of outliers follows a Bernoulli law, and the amplitude of the corrupted columns is Gaussian. In multispectral imaging for instance, this model would correspond to the presence of some spectrally different anomalies at some specific spatial positions.



(a) $\mathbf{X}_1$ versus $\mathbf{X}_2$.

(b) $\mathbf{A}^\dagger\mathbf{X}_1$ versus $\mathbf{A}^\dagger\mathbf{X}_2$.

(c) $\tilde{\mathbf{A}}^\dagger\mathbf{X}_1$ versus $\tilde{\mathbf{A}}^\dagger\mathbf{X}_2$.

Figure 1. Three sources are mixed into 4 noisy observations. Fig.(a): scatter plot of two observations, (b): scatter plot of two of the estimated sources given by $\mathbf{A}^\dagger\mathbf{X}$, (c): scatter plot of the estimated sources given by $\tilde{\mathbf{A}}^\dagger\mathbf{X}$, where $\tilde{\mathbf{A}}$ has been estimated with GMCA. The initial source contribution is represented in blue, and the one of the outliers with the red stars.

In the presence of outliers, the main challenges are summarized below:

- The MDP alone does not allow to correctly disentangle between the sources. Indeed, the largest coefficients may be related to some corrupted entries, fig.1. Thus, the largest coefficients are not the most discriminant ones, and do not permit to have a good estimation of the mixing matrix.
- There may be some mixing matrices $\tilde{\mathbf{A}}$, different from the initial $\mathbf{A}$ for which the corresponding sources (roughly given by $\tilde{\mathbf{A}}^\dagger\mathbf{X}$) are sparser than the ones obtained with $\mathbf{A}^\dagger\mathbf{X}$ fig.1c. For this reason, algorithms based on sparsity will return $\tilde{\mathbf{A}}$ and not the right mixing matrix.

Hence, in the presence of outliers, algorithms based on sparsity and more especially on the MDP such as GMCA fail to estimate correctly $\mathbf{A}$ and obviously $\mathbf{S}$ because the MDP cannot be exploited.

III. Robust Sparse BSS

In this section, we introduce the basics of a first approach for solving sparse BSS problems in the presence of outliers [19] designated as rGMCA, and discuss its main limitations.

A. A first approach: rGMCA

Structural differences between $\mathbf{O}$ *and* $\mathbf{AS}$*:* In the algorithm rGMCA, robustness to outliers is obtained by exploiting the difference of structure between the outliers and $\mathbf{AS}$. Since the sources are sparse and respect the MDP, the entries of $\mathbf{AS}$ are clustered along the directions given by the columns of $\mathbf{A}$, while the entries of $\mathbf{O}$ are distributed in general position fig.1a. Hence, by using only the sparsity and the MDP, it should be possible to determine $\mathbf{A}$ by finding these clustering directions. Moreover, if $n \ll m$, then the difference of structure between $\mathbf{AS}$ and $\mathbf{O}$ is further enhanced by the fact that the outliers are unlikely to lay exactly in the span of $\mathbf{A}$. It has been proven in [11], that even without the parsimony assumption, it is possible to separate the term $\mathbf{AS}$ from $\mathbf{O}$ by using the fact that $\mathbf{AS}$ is low-rank while $\mathbf{O}$ are randomly distributed, and hence are in general position. These points have motivated the conception of the algorithm rGMCA, described in the next paragraph.

The rGMCA algorithm: The algorithm rGMCA is based on GMCA to exploit the parsimony of the sources. It performs by estimating explicitly the mixing matrix, the sources and the outliers. To improve the robustness to outliers, the rGMCA algorithm implements a weighting scheme that penalizes the large corrupted samples in the data domain, as follows:

$$\underset{\mathbf{A},\mathbf{S},\mathbf{O}}{\text{minimize}} \frac{1}{2} \left\| \left(\mathbf{X} - \mathbf{AS} - \mathbf{O}\right) \mathbf{W}_O \right\|_2^2 + \sum_{i=1}^{n} \lambda_i \left\| \mathbf{S}_i \right\|_1 + \beta \left\| \mathbf{O} \right\|_1 , \tag{3}$$

where $\mathbf{W}_O$ the diagonal penalizing matrix whose elements are such that $\mathbf{W}_{O\,t,t} = \frac{1}{\epsilon + \|\mathbf{O}^t\|_1}$, where ϵ depends on the amplitude of the sources [19]. The role of the weighting scheme is to penalize the most corrupted samples so as to limit their influence for the estimation of $\mathbf{A}$. That is why the weights are defined according to the ℓ_1 norm of the current estimate of the outliers: such a penalization therefore increases with the amplitude of the outliers.
The rGMCA alternates, in an iterative scheme, between the estimation of $\mathbf{A}$ and $\mathbf{S}$ from $\mathbf{X} - \mathbf{O}$ and the outliers, which are estimated by estimating a sparse component from the remaining residual $\mathbf{X} - \mathbf{AS}$.

Limitations of rGMCA: The algorithm rGMCA suffers from two main limitations:

- In the presence of a large amount of outliers, whose amplitude is smaller than the sources, the value of the weighting scheme $\mathbf{W_O}$, which depends on the amplitude of the outliers, may not be large enough to correctly penalize the corrupted samples in the source domain.
- If the number of sources is close to the number of observations, rGMCA can fail to estimate $\mathbf{A}$ [19]. Indeed, in this situation, the outliers also lay in the range of $\mathbf{A}$. It then becomes much more difficult to distinguish between the contribution of the outliers from the one of the sources.

Hence, the exploitation of the sparsity by using GMCA combined with a weighting scheme based on the estimated outliers can only be used reliably if the number of observation is larger than the number of sources.

B. Discriminating between the sources and the outliers

It is natural to exploit the difference of structures between $\mathbf{AS}$ and $\mathbf{O}$ in the data domain in the spirit of the PCP algorithm [11] or rGMCA. However, when the number of observations tends to the number of sources, thus property can't be used to separate out the sources and the outliers. In the following, we propose a new approach, which allows alleviating this important limitation.

Separation of the outliers and sources in the source domain: Instead of using the difference of structure between $\mathbf{AS}$ and $\mathbf{O}$ in the data domain (by looking at $\mathbf{X}$), we propose to base our approach on the behavior of the outliers in the source domain (by observing $\mathbf{A}^\dagger\mathbf{X}$).
Since we assume that the outliers are in general position in the data domain, they should not cluster in any specific direction in the observations domain neither in any estimated source domain. This can be observed in fig.1b and fig.1c: in both cases, the outliers in the source domain (initial and estimated) are broadly distributed and non-sparse. On the other hand, the sources contribution are jointly sparse in the source domain generated by $\mathbf{A}$. Hence, the non-sparse samples are clearly non-discriminant for the separation as they correspond to corrupted samples. As we will see, this situation shares some similarities with the case where the sources to be estimated are partially correlated [20].

Similarities between the BSS problems in the presence of outliers and in the presence of partially correlated sources: In real-world applications, relevant sources can be partially correlated [20]. These correlations raise several issues which have been highlighted in [20]. Interestingly, this specific context is somehow similar to case where outliers contaminate the data:

- The MDP may be not respected: the samples shared by several sources can be the largest ones.
- There may be another matrix $\tilde{\mathbf{A}}$ such that the estimated sources $\tilde{\mathbf{A}}^\dagger\mathbf{X}$ are jointly sparsest than the initial ones.

In the determined case, the model in the presence of outliers can be recast as follows: $\mathbf{X} = \mathbf{A}\left(\mathbf{S} + \mathbf{A}^\dagger\mathbf{O}\right)$. Since $\mathbf{A}^\dagger\mathbf{O}$ is very likely to have only few active columns and to be in general position, it can be seen as corresponding to partial correlations between the sources $\mathbf{S} + \mathbf{A}^\dagger\mathbf{O}$. The main difference between the two problems is that in the over-determined case, the outliers do not necessarily lay in the span of $\mathbf{A}$: they cannot be seen as being some partially correlated entries of $\mathbf{S}$. Moreover, the behaviors of the partially correlated entries of $\mathbf{S}$ and the outliers in the source domain are exactly the same: they are not sparse whereas (the remaining of) the entries of $\mathbf{S}$ are. In [20], the authors have proposed a new method, coined AMCA, to handle these samples in the source domain. Based on the similarities between these two different contexts, we propose to build upon AMCA to design a novel robust sparse BSS algorithm.

IV. Robust AMCA Algorithm

In this section, we propose a new method to solve sparse BSS problem in the presence of outliers, which builds upon the AMCA and rGMCA algorithms. Similarly to the rGMCA algorithm [19], it estimates the mixing matrix, the sources and the outliers, and further implements a weighting scheme in the spirit of the AMCA algorithm to better discriminate between the outliers and the sources in the source domain.

A. Algorithm rAMCA

In the spirit of [19], we propose to estimate jointly $\mathbf{A}, \mathbf{S}$ and $\mathbf{O}$ by exploiting the sparsity of the sources and the outliers. This is performed by minimizing the following problem:

$$\underset{\mathbf{S},\mathbf{A},\mathbf{O}}{\text{minimize}} \frac{1}{2} \left\| (\mathbf{X} - \mathbf{AS} - \mathbf{O}) \mathbf{W} \right\|_2^2 + \sum_{i=1}^{n} \lambda_i \left\| \mathbf{S}_i \right\|_1 + \beta \left\| \mathbf{O} \right\|_{2,1} . \tag{4}$$

The $\ell_{2,1}$ norm, defined such as $\left\| \mathbf{O} \right\|_{2,1} = \sum_{j=1}^{m} \left\| \mathbf{O}^j \right\|_2$, to favor solutions $\mathbf{O}$ with few entirely active columns. This regularization term is well suited to capture outliers that are distributed in general position in the data domain.
The weight matrix $\mathbf{W}$ is a key element. Indeed, in the spirit of [20], the role of $\mathbf{W}$ is to assign each sample a weight so as to penalize those which are the most detrimental to the separation process (*i.e.* the non-sparse samples in the source domain). In the case of partially correlated sources [20], such a weighting scheme has allowed to tackle unsupervised source separation problems when the MPD cannot be exploited. In practice, these non-discriminant samples are efficiently traced by the sparsity level of the estimated source samples. In the next, for each sample t, the weight $\mathbf{W}_{t,t}$ is chosen such that:

$$\mathbf{W}_{t,t} = \frac{1}{\sqrt{\left\| \mathsf{S}^t \right\|_q + \epsilon}}, \tag{5}$$

where S denotes the normalized sources $\mathsf{S}_i = \frac{\mathbf{S}_i}{\left\| \mathbf{S}_i \right\|_2}$ and where ϵ is a scalar typically small used to avoid numerical issues. The parameter q is in the range $[0, 1]$, in order to penalize the samples with several significant sources entries, which are the most detrimental to the separation.

This problem is non-convex but can be tackled using a minimization procedure suc as the Block Coordinate Relaxation (BCS - see [21]). Instead of estimating alternatively $\mathbf{A}, \mathbf{S}$, and $\mathbf{O}$, we propose the structure presented in Alg.1: the mixture parameters $\mathbf{A}$ and $\mathbf{S}$ are estimated jointly in an inner loop for a $\mathbf{O}$, whereas when the outliers are updated from $\mathbf{X} - \mathbf{AS}$. These two steps are described in the next paragraph.

Algorithm 1: rAMCA

Input: $\mathbf{X}$

First estimation of $\mathbf{A}$ and $\mathbf{S}$:

$$\tilde{\mathbf{A}}, \tilde{\mathbf{S}} = \underset{\mathbf{S},\mathbf{A}}{\text{minimize}} \frac{1}{2} \left\| (\mathbf{X} - \mathbf{AS}) \mathbf{W} \right\|_2^2 + \sum_{i=1}^{n} \lambda_i \left\| \mathbf{S}_i \right\|_p$$

while $k < K$ **do**

Update $\tilde{\mathbf{O}}$:

$$\tilde{\mathbf{O}} = \underset{\mathbf{O}}{\text{argmin}} \frac{1}{2} \left\| \left(\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}} - \mathbf{O} \right) \mathbf{W} \right\|_2^2 + \beta \left\| \mathbf{O} \right\|_{2,1}$$

Update $\mathbf{A}$ and $\mathbf{S}$:

$$\tilde{\mathbf{A}}, \tilde{\mathbf{S}} = \underset{\mathbf{S},\mathbf{A}}{\text{minimize}} \frac{1}{2} \left\| \left(\mathbf{X} - \tilde{\mathbf{O}} - \mathbf{AS} \right) \mathbf{W} \right\|_2^2 + \sum_{i=1}^{n} \lambda_i \left\| \mathbf{S}_i \right\|_p$$

end while

return $\tilde{\mathbf{S}}$, $\tilde{\mathbf{A}}$, $\tilde{\mathbf{O}}$.

B. Estimating $\mathbf{A}$ and $\mathbf{S}$

Assuming the outliers are fixed $\mathbf{O}$, the mixing matrix and sources are estimated by solving the following minimization problem:

$$\underset{\mathbf{A},\mathbf{S}}{\text{minimize}} \frac{1}{2} \left\| (\mathbf{X} - \mathbf{O} - \mathbf{AS}) \mathbf{W} \right\|_2^2 + \sum_{i=1}^{n} \lambda_i \left\| \mathbf{S}_i \right\|_p . \tag{6}$$

The problem is tackled by estimating alternatively $\mathbf{S}$ and $\mathbf{A}$, Alg.2:

- The estimation of $\mathbf{S}$ for fixed $\mathbf{A}$ is given by:

$$\underset{\mathbf{S}}{\text{minimize}} \frac{1}{2} \left\| (\mathbf{X} - \mathbf{O} - \mathbf{AS}) \mathbf{W} \right\|_2^2 + \sum_{i=1}^{n} \lambda_i \left\| \mathbf{S}_i \right\|_1 .$$

  Unless $\mathbf{A}$ is orthogonal, the previous problem does not admit a closed form solution. The authors in [20] propose to approximate this step with a projected least-square to limit the computational cost of the update:

$$\mathbf{S}_i = \mathcal{S}_{\lambda_i}\left(\mathbf{A}^\dagger \left(\mathbf{X} - \mathbf{O} \right) \right) . \tag{7}$$

- The estimation of $\mathbf{A}$ is given by:

$$\underset{\mathbf{A}}{\text{minimize}} \frac{1}{2} \left\| (\mathbf{X} - \mathbf{O} - \mathbf{AS}) \mathbf{W} \right\|_2^2 ,$$

  whose minimizer is

$$\mathbf{A} = (\mathbf{X} - \mathbf{O}) \mathbf{W} (\mathbf{SW})^\dagger .$$

Algorithm 2: Estimation of $\mathbf{A}$ and $\mathbf{S}$

Input: $\mathbf{X} - \mathbf{O}$

Initialize $\tilde{\mathbf{A}}$ and q.

while $k < K$ **do**

Update $\tilde{\mathbf{S}}$ and $\tilde{\lambda}$

Update the weighting matrix $\tilde{\mathbf{W}}$

Update the mixing matrix $\tilde{\mathbf{A}}$

Decrease q

end while

return $\tilde{\mathbf{S}}$, $\tilde{\mathbf{A}}$.

Choice of the parameters: The choice of the regularization parameters $\left\{ \tilde{\lambda}_i \right\}_{i=1..n}$ plays a key role in sparse BSS

problems: the separation procedure is remarkably improved by employing a decreasing thresholding strategy [7]. The thresholds are chosen automatically so that an increasing number of entries are selected at every iteration. The final threshold is typically $3\sigma_i$, where σ_i stands for the standard deviation of the noise contaminating the ith source. If these values are not known, they can be estimated empirically using the median absolute deviation [7], [20].

In the following, we opt for the ℓ_1 penalization ($p = 1$) for two reasons: i) the ℓ_1 norm is more convenient to use because it is convex, and ii) it has been emphasized in [22] that the ℓ_1 norm tends to favor solutions whose corresponding sources are clustered along the axis given by the columns of the mixing matrix. This is particularly important in the presence of outliers, as only the contributions of the sources are clustering whereas the outliers are in general position.

The decreasing strategy used for the q norm in the weights of $\mathbf{W}$ is well motivated in [20]. In brief, a compromise between a strong penalization of the detrimental, non-sparse, samples and a one weaker at the beginning of the separation process is found by adopting a decreasing strategy for q. The value of the parameter q is set so that $q = \frac{0.1^{\frac{k}{K-1}}}{2}$ at the kth iteration, by starting at 0.5.

Last, the number of iterations is set to $K = 1000$, which turned to be a good compromise in the numerical experiments.

C. Estimating the outliers

In the rAMCA algorithm, the estimation of $\mathbf{O}$ given $\mathbf{A}$ and $\mathbf{S}$ is carried out by solving the problem in eq.(4):

$$\underset{\mathbf{O}}{\text{minimize}}\, \frac{1}{2} \left\| (\mathbf{X} - \mathbf{AS} - \mathbf{O})\, \mathbf{W} \right\|_2^2 + \beta \left\| \mathbf{O} \right\|_{2,1}. \qquad (8)$$

This problem admits a closed form solution:

$$\mathbf{O}^k = (\mathbf{X} - \mathbf{AS})^k \times \left(1 - \frac{\tilde{\beta}^k}{\left\| (\mathbf{X} - \mathbf{AS})^k \right\|_2} \right)_+, \qquad (9)$$

where $\mathbf{O}^k$ is the kth column of the estimated outliers and $\tilde{\beta} := \frac{\beta}{\mathbf{W}}$ is a row vector of size t, which is used to simplify the notations.

According to the above expression, the estimation of the outliers can be done quite straightforwardly from the residue $\mathbf{X} - \mathbf{AS}$. However in practice, estimating $\tilde{\mathbf{O}}$ from $\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}$ while the contribution $\tilde{\mathbf{A}}$ is not correctly estimated can dramatically hamper the separation process. This is particularly the case at the beginning of the algorithm when the mixing matrices is very likely to be far from the sought-after $\mathbf{A}$.

Detecting outliers in the source domain: The estimated residue $\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}$ can be expressed as the following:

$$\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}} = \Delta\mathbf{AS} + \mathbf{O} + \mathbf{N},$$

where $\Delta\mathbf{AS} = \mathbf{AS} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}$ corresponds to the estimation error of the contribution of the sources in the data $\mathbf{AS}$.

According to eq. (9), the estimated outliers $\tilde{\mathbf{O}}$ are proportional to this residue. This can be dominated by $\Delta\mathbf{AS}$ if the estimation error of $\tilde{\mathbf{A}}$ is large, which is very likely the case at the beginning of the separation process. For that reason, estimating the outliers from the residue $\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}$ will largely propagate and amplify estimation errors made during the separation process.

In practice, estimating iteratively the support of $\mathbf{O}$ and then its amplitude directly from $\mathbf{X}$ provides a more conservative but more efficient estimation procedure.

It is important to notice that the outliers are not sparse in the source domain (in the sense that the active columns of $\mathbf{O}$ are not significantly one sparse), while the sources are. Therefore, an efficient procedure to discriminate between the outliers and the sources consists in identifying the corrupted samples based on their sparsity level: if the columns of estimated sources $\tilde{\mathbf{S}}^k$ are not sparse enough then they can be deemed as being corrupted, and vice-versa.

A powerful measure of the sample sparsity in the *source domain* is the δ-density [23] of each column j $\delta(\tilde{\mathbf{S}}^j) = \frac{\left\| \tilde{\mathbf{S}}^j \right\|_1}{\left\| \tilde{\mathbf{S}}^j \right\|_\infty}$. This ratio takes its values between 1 (for exactly 1 sparse column) and n (for a column whose entries have a same amplitude) for non-zero column. It is thus independent of the amplitude of the columns and well suited for sparse and approximately sparse signals.

In this framework, discriminating between sparse and non-sparse source samples can be performed by defining some threshold α so that a source sample with a δ-density larger than α is very likely contaminated with an outlier.

Determining a precise numerical value for α is challenging without any precise statistical modeling of the sources and the outliers. In the next, we will make use of the following simple assumptions:

- the sources are drawn from a Laplacian law $\mathcal{L}$.
- the amplitude of the outliers in the sources domain follows a Gaussian law $\mathcal{N}$, which is well suited to model samples that are distributed in general position.

Let us notice that the variances of these statistical models do not matter since the δ density of the source samples is not sensitive to their amplitude. Then, α is chosen so that for any column vector X of size n:

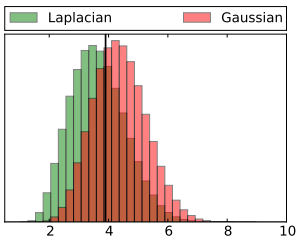
$$\Pr\left(\delta(X) = \alpha |\, X_i \sim \mathcal{L}\right) = \Pr\left(\delta(X) = \alpha |\, X_i \sim \mathcal{N}\right).$$

Similarly to classical hypothesis testing, the value of α is derived numerically by taking the value when the histogram of the δ density of the Laplacian law and the one of the Gaussian law intersect (see fig.2). In the present work, this value has been derived from simulations made with a number of samples equal to $n \times 1e^6$.

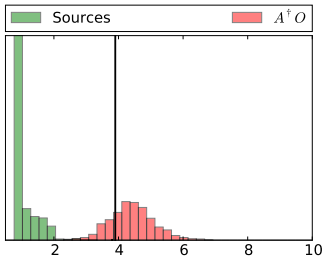
To summarize, every sample whose δ density of the sources is larger than α is considered as being corrupted and consequently is part of the support of the outliers:

$$\tilde{\mathbf{O}}^k = \begin{cases} 0 \text{ if } \delta(\tilde{\mathbf{S}}^k) < \alpha \\ \mathbf{X}^k \times \left(1 - \frac{\tilde{\beta}}{\left\| \mathbf{X}^k \right\|_2} \right) \text{ otherwise} \end{cases}$$

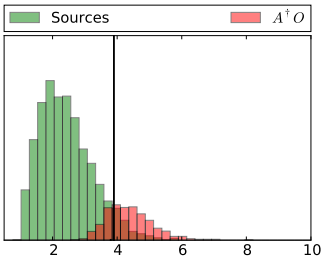
where $\tilde{\beta} = \mathrm{mad}(\mathbf{X} - \mathbf{AS} - \mathbf{O}) \times \sqrt{2} \times \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}$ corresponds to an estimation of $\mathbb{E}\left\{ \left\| \mathbf{N}^k \right\|_2 \right\}$, and thus limits the impact of the Gaussian noise.



(a) Laplacian law



(b) Bernoulli-Gaussian law



(c) Generalized Gaussian law

Figure 2. Values of the δ-density, for every non-zero samples in three examples. Fig.(a): in green, histogram of 10^6 samples drawn from a Laplacian law, and in red from a Gaussian law. Fig(b) and (c) histogram of sources drawn from a Bernoulli-Gaussian law for (b) and a generalized Gaussian law (parameter 0.3) for (c) in green, and in red, histogram of $\mathbf{A}^\dagger\mathbf{O}$ with 20 % of corruption, 10 sources, 10 observations and $t = 4096$. In the three examples, the vertical black line symbolizes the value of α for 10 sources (numerically around 3.9).

This procedure does not fully guarantee that the outliers are estimated without false detections. This may fail when the sources are very badly estimated, which is likely the case at the beginning of the separation process where $\mathbf{A}$ is far from the right solution. To limit false detections, the procedure is carried out on only a limited subset residue samples, which are composed of the entries with the largest amplitudes. These samples are more likely to be contaminated with outliers and are therefore the most detrimental to the separation process. The number of samples to be tested then grows at each iteration. This scheme is less sensitive to mis-estimations of the mixing matrix/sources during the first iterations of the rAMCA algorithm, which makes it less prone to false outlier detections. In practice, at the kth iteration in Alg.IV-A, the support to be tested for outliers is fixed to the $5k\%$ largest entries of $\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}$.

Algorithm 3: EstimationO

Input: $\mathbf{X}, \tilde{\mathbf{O}}, \tilde{\mathbf{S}}, \tilde{\mathbf{A}},$ the iteration number k.
Estimate the $5k\%$ largest entries of $\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}} - \tilde{\mathbf{O}}$.
Refine the support by keeping the samples with a δ density larger than α.
Add the previous support of the outliers
Update the amplitude of the outliers from this new support.
return $\tilde{\mathbf{O}}$.

V. Numerical Experiments

A. Experimental protocol

In this section, the performances of rAMCA are compared to the ones of the following standard BSS algorithms: GMCA, AMCA (modified version presented in the paper), PCP+GMCA [11] and the minimization of the β-divergence (implementation similar to [24]). In the first part of this section, their performances are evaluated on various scenarii with synthetic data, which allows performing Monte-Carlo simulations.

a) Data Setting: In this section, the comparisons are carried out on synthetic data in order to illustrate the impact of parameters such as the percentage of corrupted data or the number of observations with Monte Carlo simulations (48 simulations). The data are generated as follows:

- A total of 8 sources (if not otherwise stated) are drawn from a Bernoulli-Gaussian law. The sources are 5% sparse, and the standard deviation of their amplitude σ_S is 100. The number of samples t is fixed to 4096.
- The mixing matrix is drawn according to a normal law with zero mean. The columns of $\mathbf{A}$ are normalized to unit ℓ_2 norm. In the over-determined setting, the number of observations is fixed to $m = 20$, if not stated otherwise.
- The outliers are generated so as to corrupt at random a low number of data samples (*i.e.* columns of $\mathbf{X}$). The activation of these corrupted columns is drawn according to a Bernoulli process with probability ρ, which fixes the average number of corrupted columns to ρt. The amplitude of the outliers is drawn at random from a Gaussian distribution with zero mean and standard deviation σ_O.
- The noise is generated according to a Gaussian distribution with zero mean. Its standard deviation is designated by σ_N, which is set to 0.1 by default.

b) Performance criterion: We emphasized in [19] that the algorithms listed above do not all yield a precise estimation of the sources but rather provide a robust estimation of the mixing matrix. Therefore, in the remaining of this paper, we will focus on assessing the performances of these algorithms based on the mixing matrix criterion Δ_A [7]. This criterion provides a global indicator of estimation accuracy of estimation of the mixing matrix $\Delta_A = \frac{\left\|\mathbf{P}\tilde{\mathbf{A}}^\dagger\mathbf{A} - \mathbf{I}\right\|_1}{n^2}$ where the matrix $\mathbf{P}$ corrects for the permutation indeterminacy.

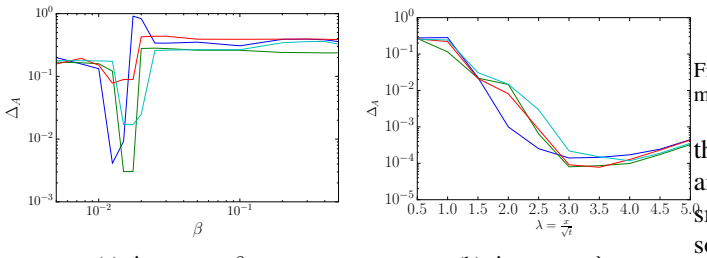
Additionally, for every simulation and for each algorithm, we record the number of runs for which $\mathbf{A}$ has been *correctly* recovered (normalized to 1). The mixing matrix is said to be correctly recovered if, for every column of $\mathbf{A}$, the angle between the estimated and true column is lower than $5°$: $\arccos(\left\langle \tilde{\mathbf{A}}^i, \mathbf{A}^i \right\rangle) < 5°$. This quantity provides a good criterion to evaluate the reliability of the algorithms.

B. Choosing the parameters of each algorithm

Comparisons are performed with the β-divergence minimization algorithm as well as the combination PCP+GMCA. These two methods requires setting up parameters, which are generally complicated to choose in practice:

- It is almost impossible to determine *a priori* the best value of the parameters since they depend on the sought-after sources, mixing matrix and the outliers, which are unknown. Fixing their value without the ground truth is therefore challenging.
- Testing the algorithms for different values of these parameters sampled on a fine grid is highly time consuming, especially if one considers that the results are sensitive to the parameter value.

To illustrate this issue, fig.3 shows the variations of Δ_A versus the value of the two parameters for 4 examples in which $n = 4$ sources are mixed into 4 and 40 observations for the minimization the β-divergence and PCP+GMCA. For both, 5% of the data are corrupted and $\sigma_O = 50$.



(a) Δ_A versus β. (b) Δ_A versus λ.

Figure 3. The variations of Δ_A versus the values of the parameters are drawn for the β-divergence on the left and PCP+GMCA for λ (eq.2) on the right for 4 examples.

The evolution of Δ_A with the algorithm parameters can be very sharped, even for small changes of their values. In order to provide fair comparisons, the parameters for PCP and the β divergence will be tuned as follows:

- The parameter λ of PCP increases from $\frac{0.5}{\sqrt{t}}$ to $\frac{10}{\sqrt{t}}$ with a step of $\frac{0.5}{\sqrt{m}}$ while the current value of Δ_A is smaller than 1.5 times the best error obtained with the previous parameter values. If the current error is larger, the step increases at $\frac{1.5}{\sqrt{t}}$.
- Similarly, the value of β increases from 0.005 to 1 with a step of 0.002 while the current Δ_A is smaller than 1.5 times the best error obtained with the previous parameter values.

C. Influence of the number of observations

We emphasized in [19] that the separation of the sources contribution and the outliers is more challenging if m is close to n. The ratio $\frac{m}{n}$ is therefore a crucial parameter in BSS, especially in the presence of outliers.

In the following experiment, we consider that the data are composed of 8 sources and m observations varying according to the values of the x-axis of fig.4. The amplitude of the outliers is fixed to $\sigma_O = 100$ for $n = m$. The amplitude ratio between the outliers and the sources contribution is kept constant. The percentage of outliers is fixed to 10%.

As shown in fig.4, rAMCA tends to be less influenced by

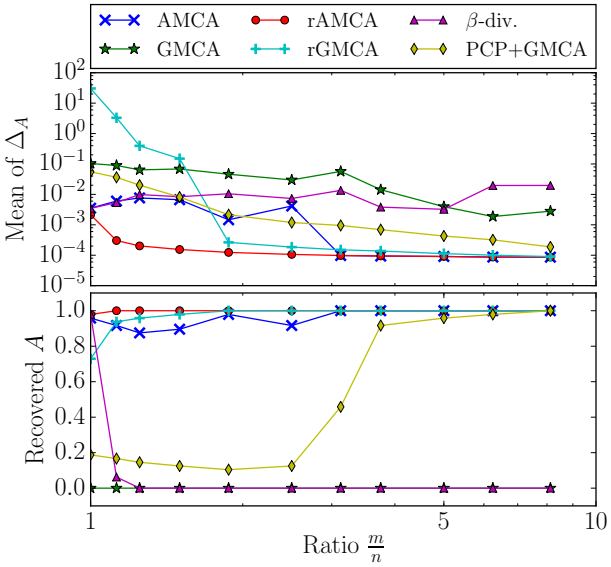


Figure 4. Influence of the number of observations on the estimation of the mixing matrix.

the number of observations. The results of all the methods are better if m is very large: the condition number of $\mathbf{A}$ is smaller and the outliers can better be distinguished from the sources contribution. Indeed, with a large m, the energy of the outliers laying in the subspace generated by $\mathbf{A}$ is lower. In this regime, the low-rankness of the term $\mathbf{AS}$ with respect to the outliers becomes a valid assumption, which makes PCP more efficient [11].

The results are not strictly improved with an increasing number of measurements for the β-divergence algorithm. It is important to notice that the β-divergence minimization algorithm has been designed for the determined case. Its application to the over-determined case requires a first dimension reduction step. This pre-processing step, which is performed by PCA, is also impacted by the presence of outliers, which hampers the performances of this algorithm when $m > n$.

In order to further illustrate the impact of the ratio $\frac{m}{n}$, the errors $\frac{\|\mathbf{S}\|_2}{\|\mathbf{S}-\hat{\mathbf{S}}\|_2}$ and $\frac{\|\mathbf{O}\|_2}{\|\mathbf{O}-\hat{\mathbf{O}}\|_2}$ are displayed for a single example. A good separation of $\mathbf{S}$ and $\mathbf{O}$ is possible if $m \gg n$ because the outliers are less likely to lay in the span of $\mathbf{A}$; this is clearly shown in fig.5. Despite an accurate recovery of $\mathbf{A}$ for rAMCA when m is small, the error made on the estimated outliers and sources is large fig.5: the separation is not possible without any additional assumption on the sources and the outliers. Moreover, these errors decrease when the ratio $\frac{m}{n}$ increases whereas the error made on $\mathbf{A}$ remains more

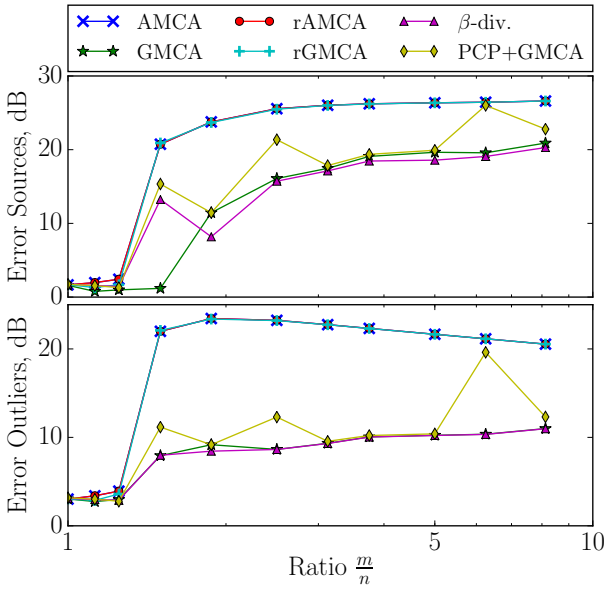


Figure 5. Influence of the number of observations on the estimations of the sources and the outliers.

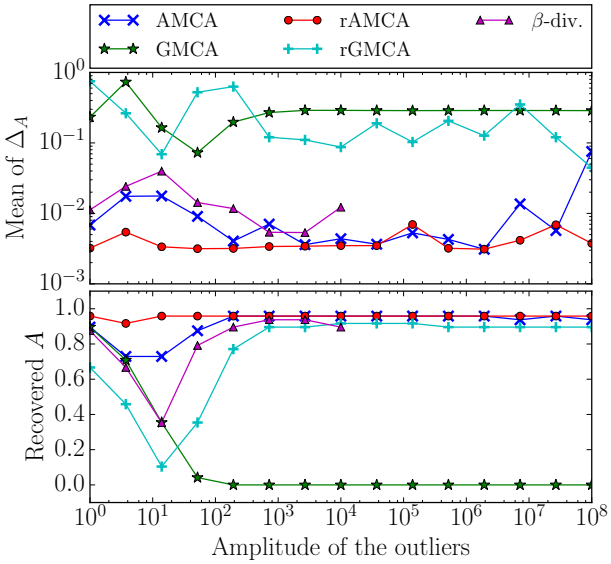


Figure 6. Influence of the amplitude of the outliers in the determined case.

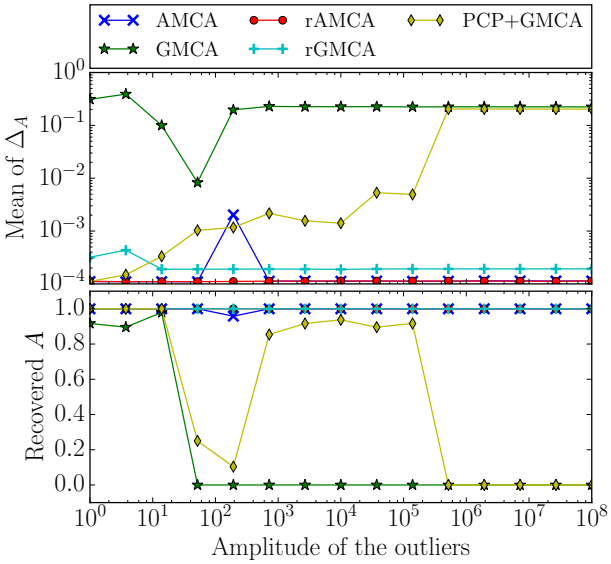


Figure 7. Influence of the amplitude of the outliers in the over-determined case.

stable: the separation benefits from enhanced estimation of $\mathbf{A}$ as well as from a lower contribution of the outliers in the range of $\mathbf{A}$.

In the following, the impact of two other parameters will be investigated: the percentage of corrupted data and their amplitudes. Because the behavior of the algorithms is quite different depending on the ratio $\frac{m}{n}$, the determined and the over-determined cases will be distinguished. In the over-determined case, the data will be composed of 8 sources and 20 noisy observations. Besides, the algorithm PCP+GMCA will only be evaluated in the over-determined case, since the low-rankness assumption makes no sense in the determined case. The β-divergence minimization methods will be studied in the determined case only, where no dimension reduction is required.

D. Influence of the amplitude of the outliers

In the following experiments, we consider that 10% of the data samples are corrupted with outliers. Fig.6 and 7 show the behavior the algorithms when the amplitude of the outliers σ_O varies.

In the determined case, the fig. 6 shows that the standard GMCA rapidly fails to correctly recover the mixing matrix when the amplitude of the outliers increases. In these experiments, the algorithms AMCA and β-divergence minimization algorithms provide very similar results. Interestingly, the rAMCA tends to be the least impacted by the amplitude of the outliers, especially when their amplitude is of the order of the source's level and when this amplitude becomes very large. When the amplitude of the outliers and the sources are close, the weighting scheme of the AMCA-based methods may be less effective at penalizing the outliers. Unlike AMCA, the rAMCA algorithm progressively removes a certain level of the outliers component, which further enhances the separation performances.

In the over-determined case, fig. 7 confirms the robustness of the rAMCA and AMCA algorithms when the amplitude of the outliers varies. In this regime, the outlier-based weighting scheme of the rGMCA algorithm makes it almost insensitive the outliers' amplitude. Finally, the PCP+GMCA algorithm seems to be highly impacted by the amplitude of the outliers since the quality of separation of the algorithm rapidly decreases.

E. Influence of the percentage of corrupted data

In this section, the amplitudes of the outliers σ_O is fixed to 100. The figures 8 and 9 show the behavior of the BSS

algorithms when the percentage of corrupted columns varies according to the values of the x-axis.

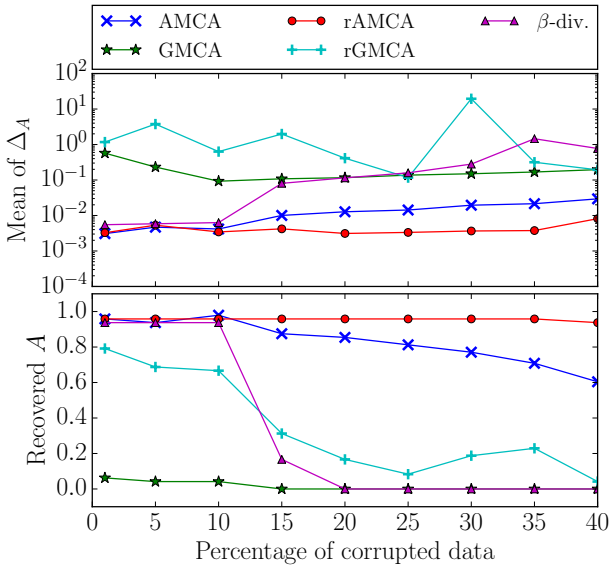


Figure 8. Influence of the number of corrupted entries in the determined case.

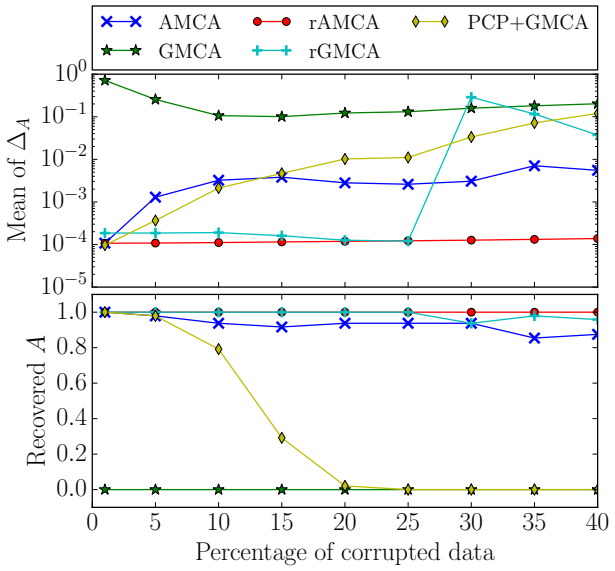


Figure 9. Influence of the number of corrupted entries in the over-determined case.

In the determined case featured in fig.8, the β-divergence algorithm is able to recover correctly the mixing matrix when the number of corrupted columns of $\mathbf{X}$ is low (*i.e.* typically below 10%). Similarly to what we highlighted previously, the rGMCA algorithm does not perform well in the determined case. On the other hand, the AMCA-based algorithms are less impacted by the number of corrupted columns. The rAMCA algorithm provides a significantly better estimate of the mixing matrix when the number of outliers is larger than 10%.

Fig. 9 displays the evolution of the mixing matrix criterion when the number of corrupted columns increases in the over-determined case. In this regime, the rGMCA, AMCA and rAMCA algorithms provides very similar results when the percentage of corrupted columns is lower than 25%. For a larger number of outliers, the performances of the rGMCA algorithms are strongly hampered in a few cases: the weighting scheme of the rGMCA algorithm tends to penalize a large number of data samples, which therefore tends to provide a larger estimation error. Though the β-divergence and PCP+GMCA algorithms provide accurate estimates of $\mathbf{A}$ when the number of corrupted samples is low, their separation performances rapidly decrease when the number of outliers increases.

It is very interesting to notice that, both in the determined and over-determined regimes, the rAMCA algorithm yields a percentage of successful estimations of $\mathbf{A}$ that is consistently close to 1. This highlights the good reliability of the rAMCA algorithm.

F. Reliability of the algorithms

Blind source separation problems are non-convex. This entails that the reliability of the algorithms to provide an accurate solution is a key characteristic. A general assessment of the reliability of the algorithms can be carried out by observing the so-called performance profiles introduced in [25]. Over T Monte-Carlo simulations, and for some scalar $\tau \geq 1$, the performance profile $P_{\mathcal{A}_i}(\tau)$ of a given algorithm $\mathcal{A}_i$ measures the fraction of simulations where the error $\Delta_{\mathbf{A}\mathcal{A}_i,r}$ made on the run r by the algorithm $\mathcal{A}_i$ is at most τ times the best error obtained by all the algorithms at this given run.

For instance, $P_{\mathcal{A}_i}(1)$ measures the percentage of simulations where the algorithm $\mathcal{A}_i$ has performed the best. More formally, the performance profile of each algorithm $\mathcal{A}_i$ is defined by:

$$P_{\mathcal{A}_i}(\tau) = \frac{\text{card}\left\{r:\ \Delta_{\mathbf{A}\mathcal{A}_i,r} \leq \tau \min_{\mathcal{A}_j:\tilde{\mathbf{A}}_{\mathcal{A}_j} \approx \mathbf{A}} \Delta_{\mathbf{A}\mathcal{A}_j,r}\right\}}{T}.$$

The performance profile therefore allows quantifying the reliability of the algorithms to accurately estimate the mixing matrix $\mathbf{A}$. In the following, the performance profiles have been computed from the $T = 1152$ simulations in the over-determined case (by combining all the results of the 48 runs for the 15 different values of the parameters in V-D and the 9 different values in V-E) and $T = 816$ for the determined case (by combining all the results of the 48 runs, for all the 9 values of the parameters in V-E and for the first 8 parameters in V-D).

The performances profiles of the tested algorithms in the determined case (*resp.* over-determined case) are displayed in the left (*resp.* right) panel of fig.10. In both cases, the AMCA and rAMCA algorithms provide a good level of reliability. Indeed, in more than 80% of the simulations these algorithms provide an estimate of $\mathbf{A}$ that exhibits a mixing matrix criterion that is at worst two times larger than the best value.

In the over-determined case, rGMCA is even more robust than AMCA, but not as accurate: its performance profile reaches the value of 1 but only for $\tau > 5$. On the other hand, it is not able to reliably estimate the mixing matrix $\mathbf{A}$ in the determined

regime. The β-divergence algorithm and PCP+GMCA provide poorer performance profiles, which indicates a lower level of reliability.
In addition providing accurates estimates of the mixing matrix, these experiments also highlight the good level of reliability of the rAMCA algorithm both in the determined and over-determined case.

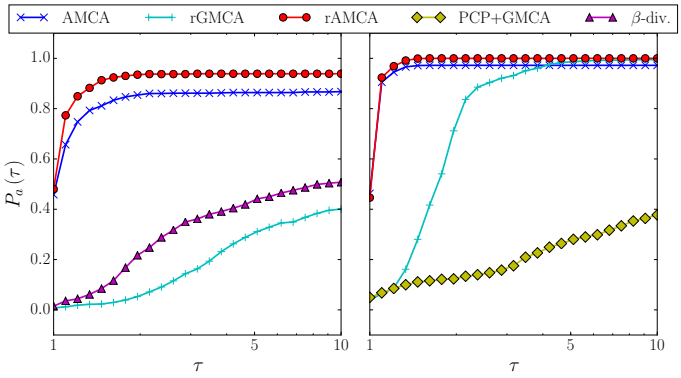
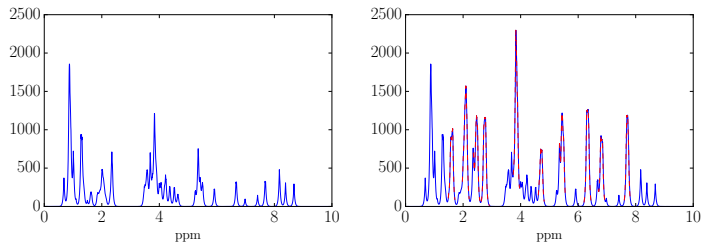


Figure 10. Performance profiles in the determined case on the left and over determined case on the right.

G. Application to NMR spectra

In this section, we propose to compare the different algorithms in a more realistic setting: the separation of Nuclear Magnetic Resonance (NMR) spectra. In the context of spectroscopy, BSS allows to identify the different molecules of the observed mixture [26]. However, the presence of instrumental artifacts is very frequent, and makes difficult the interpretation of the data. Such artifacts can be approximated by outliers contaminating entire columns of the data matrix [27], which is the case we investigate in the present article.
Following [26], the sources are composed of 6 theoretical NMR spectra of the cholesterol, folic acid, adenosine, oleic acid, menthone and saccharose extracted from the SDBS database[1]. These spectra are further convolved with a Laplacian kernel of varying width at half maximum (implementation from pyGMCA[2]), which models the resolution of the instrument. The set of corrupted data samples is fixed to 10 blocks of 20 consecutive columns. Their amplitudes are drawn according to a Gaussian law, and they are further convolved with the same kernel than the sources. The amplitude of the outliers is set so that the energy of each block of outliers corresponds to the average contribution of a source in the observations $\frac{\|\mathbf{O}\|_2}{10} = \frac{\|\mathbf{AS}\|_2}{n}$. Examples of simulated and corrupted NMR data are displayed in fig.11. In the following experiments, the data made of 10 mixtures computed with a random positive mixing matrix.
The resulting sources admit a sparser distribution in the wavelet domain. Subsequently, the data are transformed in the undecimated wavelet transform [28] prior to applying the BSS algorithms. Let us notice that the same wavelet transform is used for the outliers and the sources because they have a similar morphology in the present setting.

[1]http://sdbs.db.aist.go.jp
[2]http://www.cosmostat.org/software/gmcalab/



(a) Noiseless observation. (b) Corrupted observation.

Figure 11. Illustration of one observation $\mathbf{X}_i$, without (on the left) and with outliers (on the right, corrupted entries are represented with the dashed line).

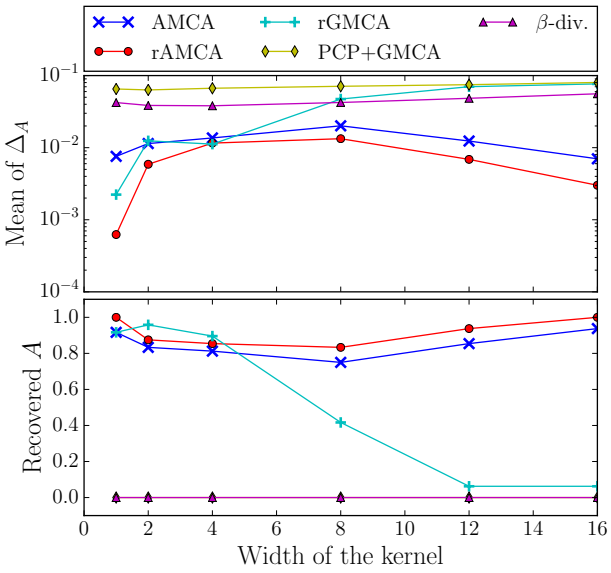


Figure 12. Performances of the different algorithms versus the width of the kernel.

In the previous experiments, we evaluated the separation performances the algorithms in the case of exactly sparse signals. The NMR sources we consider in this section rather exhibit an approximately sparse distribution in the wavelet domain. In this section, we propose to evaluate the behavior the robust BSS algorithms when both the sources and the outliers follow an approximate rather than exact sparse model. A simple way to evaluate the behavior of the algorithms with respect to the sparse model is to evaluate their performances when the width of the convolution kernel increases. Low width values will make the source model close to the exact sparse model while large values will provide approximately sparse sources.
The figure 12 displays the evolution of the mixing matrix criterion when the width of the convolution kernel varies. It is interesting to notice that the PCP+GMCA and β-divergence algorithms do not provide satisfactory separation results. As well, the rGMCA provides good separation results when the width is low but it rapidly yields incorrect results when the

width of the kernel increases. Indeed, let us recall that the outliers are also approximately sparse, which makes these separation scenarii close to the cases we investigated previously where the number of outliers is very large. This is typically the kind of settings where these methods tend to fail. The rAMCA and AMCA provide the most accurate estimates of the mixing. The discrepancy with respect to the other algorithms is particularly large when the kernel has a large width. In this regime, the level of correlation between the sources increases, a phenomenon to which the AMCA algorithm is robust [20].

VI. Conclusion

In this article, we introduce a new algorithm for tackling BSS problems in the presence of outliers, which is known to be challenging. Building upon the AMCA algorithm, the proposed rAMCA algorithm performs by estimating jointly the mixing matrix, the sources and the outliers. We emphasize that disentangling between the outliers and the sources can be efficiently tackled by exploiting the sparsity of these components as well as their different distributions in the sample domain. Indeed the sources are clustered in a subspace spanned by the columns of the mixing matrix while the outliers are assumed to be distributed in general position. In the rAMCA algorithm, these two properties are first exploited by introducing a weighting scheme to penalize the contribution of the outliers in the spirit of the AMCA algorithm. As well, we use an effective outlier estimation procedure that builds upon the difference of their distribution in the source domain. Numerical experiments have been carried out on Monte-Carlo simulations with various experimental scenarii, which show that the proposed algorithm provides a robust and reliable estimation of the mixing matrix in both the over-determined and determined cases. Future work will focus on generalizing the proposed approach to enforce the sparsity of both the sources and the outliers in a transformed domain.

References

[1] J. Bobin, F. Sureau, J.-L. Starck, A. Rassat, and P. Paykari, "Joint Planck and WMAP CMB map reconstruction," A&A, vol. 563, no. A105, 2014.

[2] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, vol. 5, no. 2, pp. 354–379, 2012.

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

[4] P. Comon, "Independent component analysis, a new concept?" Signal processing, vol. 36, no. 3, pp. 287–314, 1994.

[5] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," Environmetrics, vol. 5, no. 2, pp. 111–126, 1994.

[6] M. Zibulevsky and B. Pearlmutter, "Blind Source Separation by Sparse Decomposition in a Signal Dictionary," Neural Computation, vol. 13, no. 4, pp. 863–882, April 2001.

[7] J. Bobin, J.-L. Starck, J. Fadili, and Y. Moudden, "Sparsity and morphological diversity in blind source separation," Image Processing, IEEE Transactions on, vol. 16, no. 11, pp. 2662–2674, 2007.

[8] P. G. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," IEEE Transactions on Neural Networks, vol. 16, no. 4, 2005, 992–996.

[9] N. Gadhok and W. Kinsner, "Rotation sensitivity of independent component analysis to outliers," in Electrical and Computer Engineering, 2005. Canadian Conference on. IEEE, 2005, pp. 1437–1442.

[10] Q. Li, H. Li, Z. Lu, Q. Lu, and W. Li, "Denoising of Hyperspectral Images Employing Two-Phase Matrix Decomposition," Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, vol. 7, no. 9, pp. 3742–3754, Sept 2014.

[11] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" Journal of the ACM (JACM), vol. 58, no. 3, p. 11, 2011.

[12] H. Zhang, W. He, L. Zhang, H. Shen, and Q. Yuan, "Hyperspectral Image Restoration Using Low-Rank Matrix Recovery," Geoscience and Remote Sensing, IEEE Transactions on, vol. 52, no. 8, pp. 4729–4743, Aug 2014.

[13] M. Mihoko and S. Eguchi, "Robust blind source separation by beta divergence," Neural computation, vol. 14, no. 8, pp. 1859–1886, 2002.

[14] L. Zhang, Z. Chen, M. Zheng, and X. He, "Robust non-negative matrix factorization," Frontiers of Electrical and Electronic Engineering in China, vol. 6, no. 2, pp. 192–200, 2011.

[15] B. Shen, L. Si, R. Ji, and B. Liu, "Robust nonnegative matrix factorization via ℓ_1 norm regularization," arXiv preprint arXiv:1204.2311, 2012.

[16] C. Févotte and N. Dobigeon, "Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization," arXiv preprint arXiv:1401.5649, 2014.

[17] Y. Altmann, S. McLaughlin, and A. Hero, "Robust Linear Spectral Unmixing Using Anomaly Detection," IEEE Transactions on Computational Imaging, vol. 1, no. 2, pp. 74–85, June 2015.

[18] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in Advances in Neural Information Processing Systems 16. MIT Press, 2004.

[19] C. Chenot, J. Bobin, and J. Rapin, "Robust Sparse Blind Source Separation," Signal Processing Letters, IEEE, vol. 22, no. 11, pp. 2172–2176, 2015.

[20] J. Bobin, J. Rapin, A. Larue, and J.-L. Starck, "Sparsity and Adaptivity for the Blind Separation of Partially Correlated Sources," Signal Processing, IEEE Transactions on, vol. 63, no. 5, pp. 1199–1213, March 2015.

[21] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," Journal of optimization theory and applications, vol. 109, no. 3, pp. 475–494, 2001.

[22] J. Rapin, J. Bobin, A. Larue, and J.-L. Starck, "Sparse and Non-Negative BSS for Noisy Data," IEEE Transactions on Signal Processing, vol. 61, pp. 5620–5632, 2013.

[23] C. Studer, "Recovery of Signals with Low Density," arXiv preprint arXiv:1507.02821, 2015.

[24] N. Gadhok and W. Kinsner, "An Implementation of β-Divergence for Blind Source Separation," in Electrical and Computer Engineering, 2006. CCECE'06. Canadian Conference on. IEEE, 2006, pp. 1446–1449.

[25] E. D. Dolan and J. J. Moré, "Benchmarking optimization software with performance profiles," Mathematical programming, vol. 91, no. 2, pp. 201–213, 2002.

[26] J. Rapin, J. Bobin, A. Larue, and J.-L. Starck, "NMF with Sparse Regularizations in Transformed Domains," SIAM Journal on Imaging Sciences, vol. 7, no. 4, pp. 2020–2047, 2014.

[27] J. Rapin, A. Souloumiac, J. Bobin, A. Larue, C. Junot, M. Ouethrani, and J.-L. Starck, "Application of Non-negative Matrix Factorization to LC/MS data," Signal Processing, p. 8, 2015.

[28] J.-L.Starck, J.Fadili, and F.Murtagh, "The Undecimated Wavelet Decomposition and its Reconstruction," IEEE Transactions on Image Processing, vol. 16, pp. 297–309, 2007.