Blind separation of sparse sources in the presence of outliers

Cécile Chenot*, Jérôme Bobin

CEA, IRFU, SAp - SEDI, 91191 Gif-sur-Yvette Cedex, France

Abstract

Blind Source Separation (BSS) plays a key role to analyze multichannel data since it aims at recovering unknown underlying elementary sources from observed linear mixtures in an unsupervised way. In a large number of applications, multichannel measurements contain corrupted entries, which are highly detrimental for most BSS techniques. In this article, we introduce a new *robust* BSS technique coined robust Adaptive Morphological Component Analysis (rAMCA). Based on sparse signal modeling, it makes profit of an alternate reweighting minimization technique that yields a robust estimation of the sources and the mixing matrix simultaneously with the removal of the spurious outliers. Numerical experiments are provided that illustrate the robustness of this new algorithm with respect to aberrant outliers on a wide range of blind separation instances. In contrast to current robust BSS methods, the rAMCA algorithm is shown to perform very well when the number of observations is close or equal to the number of sources.

Keywords:

Blind source separation, sparse signal modeling, robust recovery, outliers.

*Corresponding author

Email addresses: cecile.chenot@cea.fr (Cécile Chenot), jerome.bobin@cea.fr (Jérôme Bobin)

1. Introduction

The fast development of multichannel sensors and imagers in a wide range of scientific fields mandates the development of dedicated data analysis tools. In this context, Blind Source Separation (BSS) plays a key role as it allows extracting relevant information in an unsupervised manner. Moreover, BSS has demonstrated its efficiency in numerous applications such as astrophysics [1] or hyperspectral unmixing [2] to only name two. In the setting of BSS, the data are made of m multichannel observations $\{\mathbf{X}_i\}_{i=1..m}$. Each one is assumed to be composed of the linear mixture of $n \leq m$ sources $\{\mathbf{S}_j\}_{j=1..n}$ with $t > m$ samples. This so-called linear mixture model is generally recast in the following matrix formulation:

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

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 objective of BSS is to estimate both $\mathbf{A}$ and $\mathbf{S}$ from the knowledge of the observations $\mathbf{X}$ only. Without any further assumption, BSS is a challenging matrix factorization problem that admits an infinite number of solutions. Prior information on the sources and/or the mixing matrix is required to tackle such an ill-posed problem. In brief, most BSS methods mainly differ from the prior used to describe the sources such as statistical independence in ICA (Independent Component Analysis), non-negativity in NMF (Nonnegative Matrix Factorization) or sparsity. For more details about standard BSS methods, we refer to the seminal book [3] and references therein.

It is well-known that the presence of noise hampers the performances of most BSS methods [4, 5]. Accounting for additional additive Gaussian noise is carried out by adding an extra noise term $\mathbf{N}$ to the linear mixture model: $\mathbf{X} = \mathbf{AS}+\mathbf{N}$. However, in a large number of applications, the observations are also contaminated with aberrant entries, rare and large errors, which are not correctly modeled by Gaussian noise models. More precisely, such spurious *outliers* include observed unexpected physical events or malfunctions of captors. Important examples are: i) stripping noise or impulse noise in hyperspectral data [6]), ii) cosmic ray contamination in astronomical images [7], iii) point sources emissions in astrophysical data [8] to only name a few. Beyond instrumental or physical artifacts, it has been recently advocated that sparse deviations from the linear mixture model can be approximated with outliers models in hyperspectral data [2]. Accounting for both Gaussian noise and outliers, we will further consider that the observations can be expressed as follows:

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

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. Extending the BSS framework to further dealing with outliers refers to *robust* BSS.

Robust BSS methods in the literature

In the current literature, only few BSS methods have been developed to manage the presence of outliers. These techniques can be split into three different groups:

- Robust Independent Component Analysis : In this framework, the sources are assumed to be mutually independent. One way to measure statistical independence of the estimated sources is to compute their mutual information, defined as the Kullback-Leibler (KL) divergence between the product of their marginal distributions and their joint distribution [3]. Unfortunately, the KL-divergence is highly sensitive to the presence of outliers [4]. To overcome this problem, the KL divergence is substituted in [9] with the more robust β-divergence. While such methods provide a robust estimation of the mixing matrix, it however does not perform any sources/outliers separation.

- Two-steps methods : These methods are comprised of two successive steps: i) removal of the outliers from the observations in a first step, and ii) separation of the sources from the "outliers-cleaned"data using standard non-robust BSS techniques. The first outliers denoising step is however complex to tackle. Recently, and to the best of our knowledge, the most powerful techniques are based on the PCP algorithm (Principle Component Pursuit - [10]), which have been proposed in [6, 11]. However, it is essential for the PCP algorithm to work that the contribution of the sources $\mathbf{AS}$ has low rank. This assumption is valid whenever $m \gg n$. If this assumption holds true for hyperspectral data, it is far from being the case in more general BSS problems such as in astrophysics [1].

- Component separation methods : These approaches aim at recovering simultaneously $\mathbf{A}, \mathbf{S}$ and $\mathbf{O}$. They have been essentially used in the NMF framework with optimization methods [2, 12–15] or a Bayesian approach [16]. These methods strongly rely on the positivity of both the sources and the mixing matrix, which is not necessarily a valid assumption in general settings. This is especially the case in imaging where the sources are more conveniently modeled in transformed domains where the non-negativity assumption does not hold.

Recently, we introduced a component separation method coined robust Generalized Morphological Component Analysis (rGMCA - [17]), which does not require assuming that the sources and/or the mixing matrix are non-negative. This algorithm emphasizes on the sparse modeling of the sources and outliers in the same signal representation. We showed that the rGMCA algorithm provides good separation performances in the over-determined setting ($m > n$) but fails at precisely solving robust BSS problems when the number of observations is close to the number of sources (determined case, $m = n$). This highly limits its suitability in applications where the number of available observations is of the order of the number of sources, such as in astrophysics [1].

Contribution:

In this article, we propose a novel algorithm, coined robust Adaptive Morphological Component Analysis (rAMCA), that generalizes the rGMCA algorithm [17]. It first builds upon the sparse modeling of both the outliers and the sources in the same transformed domain. Unlike the rGMCA algorithm, the proposed algorithm further relies on two novel elements: i) a refined modeling of the outliers in the source domain based on an analogy with partially correlated sources (see [18]), and ii) an improved outliers estimation procedure, described in Section 2 and 3 respectively. In Section 4, the proposed method is shown to yield a highly effective estimation of the mixing matrix. It also performs very well when the number of observations is close or equal to the number of sources; a challenging setting for which currently available robust BSS methods fail. Besides, we describe how the parameters of rAMCA can be automatically tuned.

Notations

Uppercase boldface letters denote matrices. The Moore-Penrose pseudo-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

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

Last, the operator MAD designates the median absolute deviation and Pr stands for probability.

2. Sparse BSS in the presence of outliers

2.1. Impact of outliers on sparse BSS methods

Based on sparse modeling [19], sparse BSS assumes that the sources $\{\mathbf{S}\}_{i=1..n}$ have sparsely distributed entries in some signal representation $\mathbf{\Phi}$:

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

where α_i is composed of a small number of non-zero elements (*i.e.* exactly sparse model) or few significant large-amplitude entries (*i.e.* approximately sparse model). Most natural signals verify such a sparsity property in a well-suited signal representation $\mathbf{\Phi}$ such as the wavelets or the curvelets to only name two [19]. Sparsity has been shown to largely improve the performances of BSS methods [3, 5, 20] since it allows for an enhanced discrimination between the sources to be estimated, which are assumed to verify some morphological diversity principle (MDP) [5]. According to the MDP, the large-amplitude coefficients of the sources in $\mathbf{\Phi}$, which encode their most salient morphological features, are distinct. For instance, sparse and statistically independent sources verify the MDP with high probability. This is illustrated in fig.1b: the source's samples in $\mathbf{\Phi}$ (blue dots) with the largest amplitudes are mainly clustered along the canonical axes. Based on this principle, the Generalized Morphological Component Analysis (GMCA) algorithm [5] tackles sparse BSS problems by seeking the sources that are jointly the sparsest possible. This is performed by minimizing:

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

where the first term is the data-fidelity term and the second term promotes the sparsity of the sources in the transformed domain $\mathbf{\Phi}$.

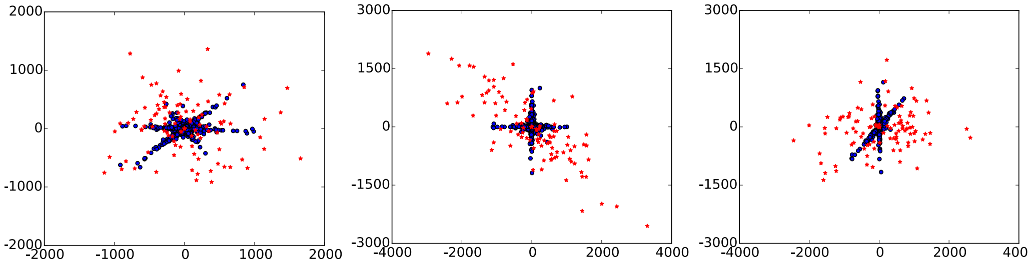
In the following, for the sake of clarity, we will assume that the sources and the outliers are sparse in the direct domain $\mathbf{\Phi} = \mathbf{I}$. Similarly to [5] and [18], all the results will be exact for any orthogonal transform $\mathbf{\Phi}$ and a good approximation for redundant sparse representations such as tight frames, which have diagonally dominant Gram matrices (*e.g.* undecimated wavelet transforms, curvelets [19], etc.).

Unlike pure sources' samples, the outliers are assumed to be distributed in general position: they do not cluster in any specific direction as illustrated in fig.1a (red dots). Since the outliers are sparse and distributed in general position, we will further assume that few columns of $\mathbf{O}$ are entirely active such as in [2].

According to the MDP, each of the sources is precisely described by its most prominent coefficients, which makes them the most informative samples to tackle the separation task. However, in the presence of outliers, the large-amplitude entries of the projected sources $\hat{\mathbf{S}} = \mathbf{A}^{\dagger}\mathbf{X}$ are more likely related to samples that are corrupted by outliers as shown in fig.1b. Consequently, when seeking for the jointly sparsest sources by unmixing the large entries of $\mathbf{X}$, sparse BSS methods are misled by the outliers. This is illustrated in the fig.1c which shows the scatter plot of the estimated projected sources $\tilde{\mathbf{S}} = \tilde{\mathbf{A}}^{\dagger}\mathbf{X}$ where $\tilde{\mathbf{A}}$ is the mixing matrix returned by GMCA: the corresponding sources are jointly sparser than the ones in fig.1b. However such a solution yields a poorly estimated mixing matrix $\tilde{\mathbf{A}}$. It highlights that applying sparse BSS method will very likely yield an erroneous solution. *By nature, sparse BSS methods are highly sensitive to the presence of outliers.*

2.2. An analogy with partially correlated sources

In [17], it clearly appears that solving robust BSS problems requires discriminating between the outliers and the sources. Since both the outliers and the sources are assumed to be sparsely represented in the same domain



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

(b) $\hat{\mathbf{S}}_1$ versus $\hat{\mathbf{S}}_2$.

(c) $\tilde{\mathbf{S}}_1$ versus $\tilde{\mathbf{S}}_2$.

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

$\mathbf{\Phi}$, sparsity alone cannot be the right separation criterion. Fortunately, both components are assumed to have different distributions: sources samples tend to cluster along the canonical axes in the source domain while the samples of the projected outliers $\hat{\mathbf{O}} = \mathbf{A}^{\dagger}\mathbf{O}$ (*i.e.* projection of $\mathbf{O}$ in the source domain) do not have any preferred clustering direction. This is testified by the difference between the distributions in the source domain of the sources samples (blue dots) and corrupted samples (red stars) in fig.1b.

If the mixing matrix $\mathbf{A}$ were perfectly known, the sources would be approximated by projecting the corrupted data onto the span of $\mathbf{A}$: $\hat{\mathbf{S}} = \mathbf{A}^{\dagger}\mathbf{X}$. The estimated sources are the linear combination of the clean sources and the projected outliers: $\hat{\mathbf{S}} = \mathbf{S} + \hat{\mathbf{O}}$. Due to the projected outliers contribution which is broadly distributed, some of the largest entries of $\hat{\mathbf{S}}$ are active simultaneously in several sources (*c.f.* the red contribution in fig.1b). These shared active samples are reminiscent of the partial correlations of the sources we discussed in [18]. Indeed, the samples of partially correlated sources can be similarly divided into two groups: the discriminant samples respecting the

MDP (the jointly sparse contribution in blue in fig.1b) and the samples corresponding to the partial correlations which active simultaneously in several sources (the broadly distributed contribution in red in fig.1b). Unlike the rGMCA algorithm we introduced in [17], we propose to exploit the analogy between the impact of the projected outliers and sparse and partially correlated sources, which yields a novel robust BSS algorithm that is described in the next section.

3. Robust AMCA Algorithm

Following the analogy between the impact of outliers and partial correlations, the rAMCA algorithm will be built upon the AMCA algorithm (Adaptive Morphological Component Analysis), which has been designed to deal with partially correlated sources [18]. In this specific context, we underlined in [18] that the ability to identify correlated entries is critical to perform the separation. For that purpose, the AMCA algorithm builds upon an adaptive weighting scheme that assigns to each column of the observation coefficients $\mathbf{X}$ a weight, whose goal is to penalize correlated entries in the separation process. Details about the weighting procedure will be given below. According to [18], in the setting of partially correlated sources, the AMCA algorithm performs by minimizing the following problem:

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

where $\mathbf{W} \in \mathbf{R}^{t \times t}$ is the weight matrix.

In the spirit of [17], we propose to estimate jointly $\mathbf{A}, \mathbf{S}$ and $\mathbf{O}$ by exploiting the sparsity of the sources and the outliers. Unlike the rGMCA algorithm, we propose to further employ a weighting scheme similar to AMCA. This can be done by substituting the problem in eq. 2 with the following one:

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

The $\ell_{2,1}$ norm, defined such as $\left\|\mathbf{O}\right\|_{2,1} = \sum_{j=1}^{m} \|\mathbf{O}^j\|_2$, favors 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. This problem is non-convex but can be tackled using a minimization procedure such as the Block Coordinate Relaxation [21] (BCR). This minimization technique amounts to sequentially minimizing subsets of variables. A natural choice would consist in estimating alternatingly the three variables of interest $\mathbf{A}, \mathbf{S}$, and $\mathbf{O}$. However, we found that this choice performs poorly in practice since errors are more likely to propagate from one variable to the other during the sequence of minimization steps. We rather opted for sequential minimization of two blocks of variables: i) the couple $(\mathbf{A}, \mathbf{S})$ and ii) the outliers matrix $\mathbf{O}$. The major advantage of this choice is that it provides a much more robust estimation of $(\mathbf{A}, \mathbf{S})$ since both parameters are updated jointly and not independently from the residual $\mathbf{X} - \mathbf{O}$.

Procedure 1 rAMCA Algorithm

```
procedure rAMCA($\mathbf{X}$, $n$)
    Initialize $\tilde{\mathbf{A}}^{(0)}$ (randomly or with a PCA), $\tilde{\mathbf{S}}^{(0)} = 0$ and $\tilde{\mathbf{O}}^{(0)} = 0$.
    while $k < K$ do
        Set $\tilde{\mathbf{S}}^{(0,k)} \leftarrow \tilde{\mathbf{S}}^{(k-1)}$ and $\tilde{\mathbf{A}}^{(0,k)} \leftarrow \tilde{\mathbf{A}}^{(k-1)}$
        while $i < I$ do                                ▷ Joint estimation of $\mathbf{A}$ and $\mathbf{S}$
            Update $\tilde{\mathbf{S}}^{(i,k)}$ from (5)
            Update $\tilde{\mathbf{W}}$ from (7)
            Update $\tilde{\mathbf{A}}^{(i,k)}$ from (6)
        Set $\tilde{\mathbf{S}}^{(k)} \leftarrow \tilde{\mathbf{S}}^{(i-1,k)}$ and $\tilde{\mathbf{A}}^{(k)} \leftarrow \tilde{\mathbf{A}}^{(i-1,k)}$
        Update $\tilde{\mathbf{O}}^{(k)}$ from (9)                      ▷ Estimation of $\mathbf{O}$
    return $\tilde{\mathbf{S}}^{(k-1)}$, $\tilde{\mathbf{A}}^{(k-1)}$, $\tilde{\mathbf{O}}^{(k-1)}$.
```

3.1. Estimating the sources and the mixing matrix

Applying the BCR technique to estimate the mixing matrix and the sources amounts to minimizing the problem in Eq. 3 assuming **O** is fixed:

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

The problem shares similarities with the problem solved by the AMCA algorithm (see (2)) with the exception that it applies to the residual $\mathbf{X} - \mathbf{O}$ rather than the raw observations **X**. Following the AMCA algorithm, the problem in (4) is tackled by minimizing alternately the cost function with respect to **A** and **S** with the two following steps:

- Updating **S** assuming **A** is fixed : Minimizing (4) with respect to **S** consists in solving the following convex problem:

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

Unless **A** is orthogonal, the previous problem does not admit a closed form solution. In the spirit of alternated least-square minimization techniques, we proposed in [18] to rather approximate this step with a projected least-square, which highly limits the computational cost of the update:

$$\mathbf{S}_i = \mathcal{S}_{\lambda_i}\big(\big[\mathbf{A}^\dagger \left(\mathbf{X} - \mathbf{O}\right)\big]_i\big) \,. \tag{5}$$

- Updating **A** assuming **S** is fixed : Minimizing (4) with respect to **A** amounts to solving the following convex problem:

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

  which admits a closed form solution:

  $$\mathbf{A} = \left(\mathbf{X} - \mathbf{O}\right) \mathbf{W} \left(\mathbf{SW}\right)^\dagger . \tag{6}$$

In practice, to avoid the balance indeterminacy between **A** and **S**, we assume that the columns of **A** are normalized for the ℓ_2 norm. Similarly to what is done with AMCA, this additional constraint is handled by normalizing the columns of **A** after the projected least-squares.
Similarly to the AMCA algorithm [18], the weights play a central role. In the setting of robust BSS, they help providing robustness to the remaining outliers contribution in the estimated residual $\mathbf{X} - \mathbf{O}$. Following the analogy with partial correlations, the weights aim at penalizing entries of the estimated sources which are in general position rather than clustered along a canonical axis. The former are more likely related to residuals of outliers

while the latter are characteristics of the sources. Following [18], samples in general position can be traced by measuring the sparsity level of the columns of the estimated sources using a ℓ_q norm. Therefore, the diagonal elements of the weight matrix $\mathbf{W}$ are defined as follows:

$$\mathbf{W}_{t,t} = \frac{1}{\sqrt{\left\|\mathsf{S}^t\right\|_q + \epsilon}}, \tag{7}$$

where S denotes the normalized sources $\mathsf{S}_i = \frac{\mathbf{S}_i}{\|\mathbf{S}_i\|_2}$ and where ϵ is a scalar typically small used to avoid numerical issues. The parameter q is chosen in the range $[0, 1]$. For more details, we refer the reader to [18].

3.2. 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 (3):

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

Given that only the diagonal terms of $\mathbf{W}$ are non-zero, this problem is separable. It amounts to solve for each sample $k \in \{1..t\}$:

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

This problem is equivalent to:

$$\text{minimize}_{\mathbf{O}^k} \frac{(\mathbf{W}_k^k)^2}{2} \left\| (\mathbf{X} - \mathbf{AS})^k - \mathbf{O}^k \right\|_2^2 + \beta \left\|\mathbf{O}^k\right\|_2.$$

Then, by setting $\tilde{\beta} = \frac{\beta}{(\mathbf{W}_k^k)^2}$, we end up with:

$$\text{minimize}_{\mathbf{O}^k} \frac{1}{2} \left\| (\mathbf{X} - \mathbf{AS})^k - \mathbf{O}^k \right\|_2^2 + \tilde{\beta} \left\| \mathbf{O}^k \right\|_2 .$$

This problem has a closed form solution which has been derived in [22]:

$$\mathbf{O}^k = (\mathbf{X} - \mathbf{AS})^k \times \left(1 - \frac{\tilde{\beta}^k}{\|(\mathbf{X} - \mathbf{AS})^k\|_2} \right)_+ . \qquad (8)$$

Detecting the outliers. Most sparsity-based thresholding procedures can be interpreted as detection procedures: detecting sparse samples out of dense noise. In that case, it is customary to fix the value of the threshold based on the noise statistics [19]; we will see later that this is exactly how the thresholds $\{\lambda_i\}_{i=1,\cdots,n}$ are fixed. Similarly, and according to (8), the support (*i.e.* the set of active columns) of $\mathbf{O}$ is defined by the set of columns whose ℓ_2 norm exceeds the threshold $\tilde{\beta}$. Ideally, the columns having a ℓ_2 norm smaller than $\tilde{\beta}$ should correspond to the remaining Gaussian noise. Consequently, the values of $\tilde{\beta}$ should also be fixed based on the Gaussian noise statistics. In that case, only the Gaussian noise contributes to the residual outside the support of $\mathbf{O}$. Therefore, the samples $\left\{ \left\| (\mathbf{X} - \mathbf{AS})^k \right\|_2 \right\}_{k : \left\| \mathbf{O}^k \right\|_2 = 0}$ follow a χ law with m degrees of freedom. The value of $\tilde{\beta}$ can then be chosen based on the expected value of the χ law: $\sigma \times \sqrt{2} \times \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}$, where σ corresponds to the standard deviation of $\mathbf{N}$.

Nevertheless, relying on the noise statistics only provides a detection procedure that is not reliable in the determined case. Indeed, even if $\mathbf{A}$ is correctly recovered, the outliers are very likely to leak into the estimated sources $\tilde{\mathbf{S}}$ since they also lie in the span of $\mathbf{A}$: $\tilde{\mathbf{S}} = \mathbf{S} + \mathbf{A}^{\dagger}\mathbf{O}$, such that $\mathbf{A}\tilde{\mathbf{S}} = \mathbf{A}\mathbf{S} + \mathbf{O}$. An accurate detection of the outliers based on the residual $\mathbf{X} - \mathbf{A}\tilde{\mathbf{S}}$ is then not possible. To overcome this issue, we propose to rather build the detection procedure on a quantity that allows discriminating between the outliers and the sources, especially in the determined case.

We emphasized in Section 2.2 that in the source domain the entries of $\mathbf{S}$ are jointly sparse, *i.e.* clustered along the canonical axes, whereas the projected outliers behave as correlated non-sparse entries. In this context, the δ-density, which has been introduced in [23], provides a convenient measure of sample sparsity that allows discriminating between sparse and non-sparse columns of $\tilde{\mathbf{S}}$. The δ-density of any jth non-zero sample of the estimated sources is defined as $\delta(\tilde{\mathbf{S}}^j) = \frac{\|\tilde{\mathbf{S}}^j\|_1}{\|\tilde{\mathbf{S}}^j\|_\infty}$. This quantity takes its values between 1 (for one active entry) and n (for a column whose entries have the same amplitudes). More interestingly, it is independent of the amplitude of the columns and well suited for sparse and approximately sparse signals. In this framework, detecting the support of $\mathbf{O}$ can be performed by identifying the columns of the estimated $\tilde{\mathbf{S}}$ whose δ-density is larger than a certain threshold α that needs to be determined. This is somehow reminiscent of the outliers detection discussed in [24].

In the general setting, determining an optimal numerical value for α is challenging without an accurate statistical modeling of the sources and the outliers. In the following, we propose to use the following statistical modeling:

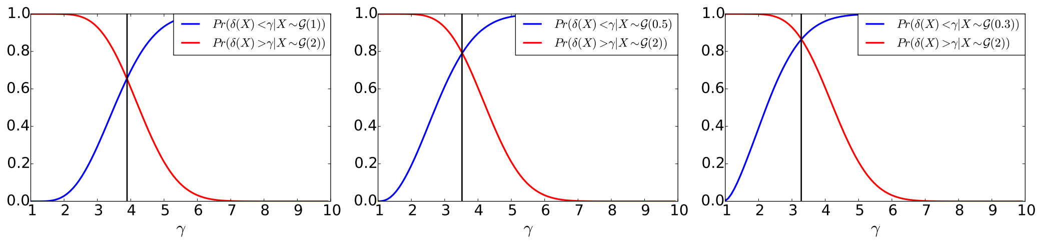
- the sources are drawn from a Generalized Gaussian law with parameter ρ denoted by $\mathcal{G}(\rho)$.

- the amplitude of the outliers in the sources domain follows a Gaussian law $\mathcal{N}$, 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 is independent of the amplitude. From this statistical model, the threshold α is derived from a classical hypothesis testing procedure such that, for any random variable X of size n:

$$\Pr\left(\delta(X) < \alpha \middle| X \sim \mathcal{G}(\rho)\right) = \Pr\left(\delta(X) > \alpha \middle| X \sim \mathcal{N}\right).$$

where $\Pr\left(\delta(X) < \alpha \middle| X \sim \mathcal{G}(\rho)\right)$ stands for the probability for the δ-density to be smaller than α assuming that every entry of X is distributed according to a Generalized Gaussian law with parameter ρ. Figure 2 illustrates three different cases with $n = 10$: (a) the case $\rho = 1$, which corresponds to a low sparsity level, and then (b) and (c), the cases $\rho = 0.5$ and $\rho = 0.3$ that correspond to realistic sparsity levels for the coefficients of sparse representations of natural signals. The value of α varies from 3.9 to 3.3. Since we have no precise prior knowledge about the distributions, we derive numerically the value α for the corresponding n from the Laplacian law (the largest possible for sparse sources respecting the MDP). This choice is quite conservative for the sources since only the samples having a δ-density larger than α are estimated as being corrupted.



(a) $\rho = 1.$

(b) $\rho = 0.3$

(c) $\rho = 0.1$

Figure 2: Numerical approximations of the cumulative distribution functions of $\delta(.)$ for different values of ρ and $n = 10$. In blue: fig.(a): $Pr(\delta(X) < \gamma | X \sim \mathcal{G}(1))$, fig.(b) $Pr(\delta(X) < \gamma | X \sim \mathcal{G}(0.5))$, fig.(c) $Pr(\delta(X) < \gamma | X \sim \mathcal{G}(0.3))$. In red: $Pr(\delta(X) > \gamma | X \sim \mathcal{N})$.

According to (8), the amplitude of the detected outliers is derived from the estimated residual $\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}$. Previously, we underline that $\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}$ is very likely to contain some errors. A more conservative but more effective choice consists in deriving the amplitude of the detected outliers from the data $\mathbf{X}$. As a summary, the outliers $\mathbf{O}$ are estimated as follows:

$$\tilde{\mathbf{O}}^k = \begin{cases} 0 \text{ if } \delta(\tilde{\mathbf{S}}^k) < \alpha \\ \mathbf{X}^k \times \left(1 - \dfrac{\tilde{\beta}}{\|\mathbf{X}^k\|_2}\right) \text{ otherwise,} \end{cases} \tag{9}$$

where $\tilde{\beta} = \text{MAD}(\mathbf{X} - \mathbf{AS} - \mathbf{O}) \times \sqrt{2} \times \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}$ and $\text{MAD}(\mathbf{X} - \mathbf{AS} - \mathbf{O})$ corresponds to a good estimate of the standard deviation of $\mathbf{N}$ if it is not known. Despite the simplicity of the statistical model used to derive a value for α and consequently β, the proposed scheme has been proved to be robust in the various numerical experiments of Section 4. Furthermore, at each iteration of the rAMCA algorithm 1, the couple $(\mathbf{A}, \mathbf{S})$ is fully re-estimated, which also makes the algorithm less sensitive to mis-estimations of the outliers $\mathbf{O}$.

3.3. Choice of the parameters

Strategy for λ: The major parameters of the sparse source separation problems are the thresholds $\left\{\tilde{\lambda_i}\right\}_{i=1..n}$. Similar to [18], we use the decreasing thresholding strategy proposed in [5] which has two interesting properties: i) it prevents the incorporation of noise in the source estimates and ii) it makes AMCA less prone to be being trapped in local minima.

Robustness to Gaussian noise: The soft-thresholding operator $\mathcal{S}_\lambda(.)$ rejects the entries having an amplitude smaller than λ. The final threshold is thus chosen based on the level of noise which contaminates the projected sources $\mathbf{A}^\dagger(\mathbf{X} - \mathbf{O}) = \mathbf{S} + \mathbf{A}^\dagger\mathbf{N}$, so as to remove the additional noise. Indeed, Gaussian noise removal ($\mathbf{A}^\dagger\mathbf{N}$) from sparse signals ($\mathbf{S}$) based on thresholding can be interpreted as a standard hypothesis testing [19]. The value of λ_i is typically set to $3\sigma_i$, where σ_i stands for the standard deviation of the noise contaminating the ith projected source, *i.e.* $(\mathbf{A}^\dagger\mathbf{N})_i$. If these values are not known, they can be estimated empirically using the Median Absolute Deviation (MAD) since MAD $\left(\left(\mathbf{A}^\dagger(\mathbf{X} - \mathbf{O})\right)_i\right) \approx$ MAD $\left((\mathbf{A}^\dagger\mathbf{N})_i\right)$ for sparse sources - see [5, 18].

Robustness to local minima: Following [5], the use of a decreasing thresholding strategy remarkably improves the separation performances since it provides more robustness to the spurious local minima. During the unmixing process, the thresholds are chosen automatically so that the number of non-zero entries of the sources is increased by a fixed amount at every iteration. More precisely, given the total number of iterations I, the jth projected source $\tilde{\mathbf{S}}_j^{(i,k)} = \left(\tilde{\mathbf{A}}^{(i-1,k)\dagger}(\mathbf{X} - \tilde{\mathbf{O}}^{(k)})\right)_j$ is thresholded at the ith iteration by:

$$\lambda_j = \text{pct}\left(|\tilde{\mathbf{S}}_j^{(i,k)}|_{|\tilde{\mathbf{S}}_j^{(i,k)}|>3\sigma_j}, 100 \times \frac{I-i}{I}\right),$$

where $\text{pct}(\mathbf{x}, v)$ denotes the vth percentile of the entries of $\mathbf{x}$.

Number of inner loops I: The number of iterations is set to I = 1000, which turned to be a good compromise in the numerical experiments.

Strategy for β: In the spirit of the decreasing value strategy used for $\tilde{\lambda}$ in AMCA, the number of eligible active samples of the estimated outliers is increased during the algorithm. More precisely, at the kth iteration, we select the outliers among the $5k\%$ largest entries of the residue in order to limit the number of false estimations. We underline that these parameters are also automatically determined: α depends only on the number of sources and β on the number of observations.

Number of outer loops K: Last, the number of outer loops is maximally set to 100. In practice, the algorithm is stopped when $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{O}}$ are jointly stabilized fig.3. More precisely, rAMCA stops at the kth outer loop if: $\max_{j=1..n}\langle\tilde{\mathbf{A}}^{(k-1)j}, \tilde{\mathbf{A}}^{(k)j}\rangle < 5^\circ$, and $\text{supp}\left(\tilde{\mathbf{O}}^{(k-1)}\right) = \text{supp}\left(\tilde{\mathbf{O}}^{(k)}\right)$, where $\text{supp}(\mathbf{x})$ denotes the support of the vector $\mathbf{x}$.

3.4. Stability of rAMCA

Since the problem (3) is not convex, we can only expect to converge to a local minima. Besides, given that the proposed strategy uses varying parameters, the convergence to a critical point, strictly speaking, cannot be proved. However, the stability of the two variables of interest, the support of the corrupted samples and the mixing matrix, is heuristically well motivated.

We propose to minimize the function using the Block Coordinate Descent (BCD) method [21]. It has been shown in [21] that minimizing (3) alternately for each variable with fixed parameters converges to a stationary point. However, in practice, minimizing (3) with the cyclic rule and with fixed parameters performs poorly: this minimizing scheme is very likely to be prone to being trapped in local minima. That is why, we minimize 3 using a sequential minimization alternating between the blocks $(\mathbf{A}, \mathbf{S})$ and $\mathbf{O}$, as well as the decreasing parameters strategy.

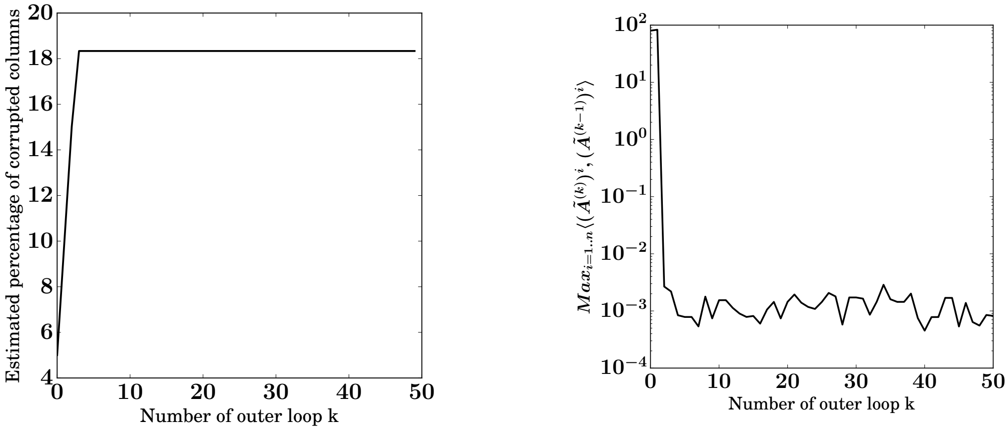
Once the detrimental outliers (or the data estimated as being detrimental) have been removed from the observations, the AMCA algorithm, whose stability has been discussed in [18], returns a similar $\mathbf{A}$ from one iteration to another (since the input $\mathbf{X} - \mathbf{O}$ is constant from one outer iteration to another one), fig.3a, 3b.

For illustrative purpose, we display the maximal angle made between the columns of $\tilde{\mathbf{A}}^{(k)}$ and $\tilde{\mathbf{A}}^{(k+1)}$ (see Section 4 for the metrics) as well as the percentage of estimated corrupting columns for $n = m = 10$ sources generated according to Section 4.1 and 30% of corrupting columns. After few outer loops, the number of estimated columns fig.3a and $\tilde{\mathbf{A}}$ almost not vary fig.3b (a variation with the maximal order of magnitude of 10^{-3} is observed for $\mathbf{A}$ due to the projected least squares).

4. Numerical Experiments

4.1. Experimental protocol

In this section, rAMCA is compared with various robust BSS algorithms:



(a) Percentage of estimated number of corrupted columns.

(b) Variation of **A** in degree across the outer loops.

Figure 3: Convergence of rAMCA.

- GMCA [5]: this standard sparse BSS method is used to illustrate the sensitivity of the *non-robust* BSS algorithms to the presence of outliers.

- AMCA [18] whose performances show the benefits of the weighting scheme (difference between AMCA and GMCA) and of the explicit estimation of **O** (difference between AMCA and rAMCA).

- rGMCA [17]: the discrepancy between its performances and the ones of rAMCA illustrates the key roles of the novel penalization and outliers detection procedure, which are, unlike rGMCA, based on the refined modeling of the outliers in the source domain.

- the robust minimization of the β-divergence [9], (implementation similar to [25]), which assumes that $m = n$ and only estimates the mixing matrix.

- the robust combination PCP+GMCA: the outliers are first estimated with PCP [10] which assumes that $m \gg n$, and then the sources and mixing matrix are estimated with GMCA.

The parameters of PCP+GMCA and of the minimization of the β-divergence are manually tuned. In the first part of this section, their performances are evaluated on various scenarios with synthetic data, which allows performing Monte-Carlo simulations.

Complexity of the different methods. The AMCA and GMCA based methods have a complexity of $\mathcal{O}\left(mnt\right)$, while PCP has a complexity given by $\mathcal{O}\left(m^2 t\right)$. Their complexities are thus similar except if $m \gg n$, what is required by PCP (low-rank assumption). The minimization of the β-divergence depends on the algorithm used to perform the minimization.

Nonetheless, we point out that in practice, if the dimensions are moderate, running rAMCA may require more computational time than running once the combination PCP+GMCA. On the other hand, PCP is a parametric method whose parameter tuning requires several trials, what is then even more time consuming.

Performance criteria. We emphasized in [17] that the algorithms listed above do not all yield a precise estimation of the sources but rather provide a robust estimation of $\mathbf{A}$. Therefore, we will focus on assessing the performances of these algorithms with respect to the mixing matrix. More precisely, we propose to evaluate the accuracy of the different algorithms as well as their reliability, which is particularly relevant since BSS problems are non-convex. The quantity $\Delta_A = \frac{\left\|\mathbf{P}\tilde{\mathbf{A}}^{\dagger}\mathbf{A}-\mathbf{I}\right\|_1}{n^2}$ is used as a global indicator of the mixing matrix estimation accuracy [5], 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 ith column is smaller than 5°: $\arccos(\left\langle \tilde{\mathbf{A}}^i, \mathbf{A}^i \right\rangle) < 5^\circ$. This quantity provides a good indicator of the reliability of the algorithms.

Data Setting. The comparisons are first 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 (unless otherwise stated) are drawn from a Bernoulli-Gaussian law whose activation rate is fixed to 5%, and the standard deviation of their amplitude σ_S to 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.

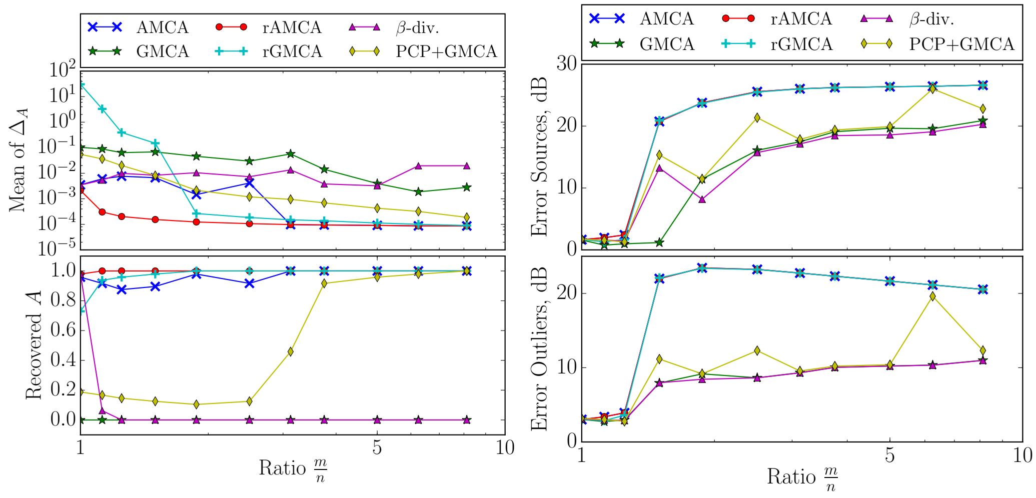
- The outliers are generated so as to corrupt at random a low number of columns of $\mathbf{X}$. The activation of these columns is drawn according to a Bernoulli process with probability ρ, which fixes the average number of corrupted columns to ρt. The amplitude of the outliers is drawn at random from 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 set to 0.1.

4.2. Influence of the number of observations

We emphasized in [17] 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 this paragraph, the data are composed of m observations. The amplitude of the outliers is fixed to $\sigma_O = 100$ for $n = m$ and then the amplitude ratio between the outliers and the sources contribution is kept constant. The percentage of outliers is fixed to 10% with $\rho = 0.1$.

As shown in fig.4a, rAMCA tends to be less influenced by the number of observations. The results of all the methods (except the β-divergence minimization algorithm) are better if m is very large: the condition number of



(a) Error made on $\tilde{\mathbf{A}}$ versus $\frac{m}{n}$.

(b) Errors made on $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{O}}$ versus $\frac{m}{n}$.

Figure 4: Influence of the number of observations on the estimations of $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{O}$.

$\mathbf{A}$ is smaller and the outliers can be better distinguished from the sources contribution since the energy of the outliers lying in the subspace generated by $\mathbf{A}$ is lower when m is large. In this regime, the low-rankness of the term $\mathbf{AS}$ becomes a valid assumption, which makes PCP more efficient [10].

The results are not strictly improved with an increasing number of measurements for the β-divergence algorithm. Since the β-divergence minimization algorithm has been designed for the determined case only, 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 and hampers the performances of this algorithm.

In order to further illustrate the impact of the ratio $\frac{m}{n}$, the errors $\frac{\|\mathbf{S}\|_2}{\left\|\mathbf{S}-\tilde{\mathbf{S}}\right\|_2}$ and $\frac{\|\mathbf{O}\|_2}{\left\|\mathbf{O}-\tilde{\mathbf{O}}\right\|_2}$ are displayed for a single example. Since the minimization of the β-divergence does not explicitly return $\mathbf{O}$ and $\mathbf{S}$, we (re)-estimate $\mathbf{O}$ and $\mathbf{S}$ by minimizing (3) for fixed $\mathbf{A}$, the mixing matrices estimated by the different algorithms. A good separation of $\mathbf{S}$ and $\mathbf{O}$ is possible if $m \gg n$ because the outliers are less likely to lie in the span of $\mathbf{A}$; this is clearly shown in fig.4b. 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.4b: 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 stable: the separation benefits from enhanced estimation of $\mathbf{A}$ as well as from a lower contribution of the outliers in the range of $\mathbf{A}$.

4.3. Influence of the number of samples t

In the following experiment, we investigate the influence of the number of samples. In order to observe the impact of this data dimension on the combination PCP+GMCA, we consider that 6 sources are mixed into 30 observations (the low-rank assumption is valid), which are corrupted by $\rho_O = 10\%$ of active outliers with $\sigma_O = 50$. We set σ_N to 0.1. The number of samples t varies according to the x-axis of fig.5.

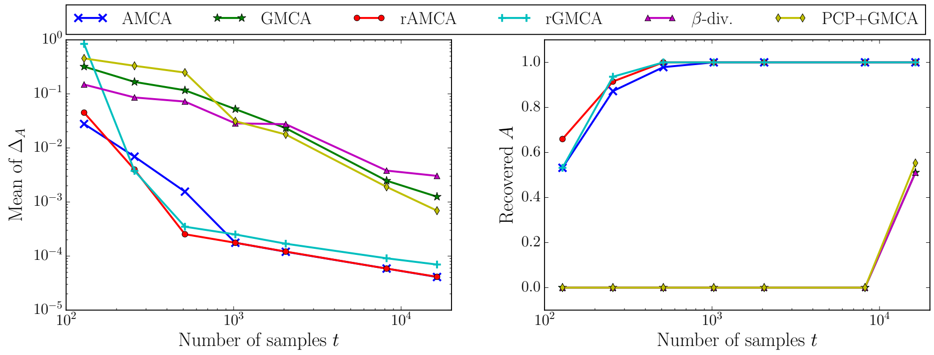


Figure 5: Performance results of the methods versus the number of samples t.

As shown in fig.5, all the algorithms are less reliable if only few samples are available since the clustering aspect of the sources contributions is not significant (an unmixing, even without outliers, is challenging if only few samples are available). Besides, all the strategies become more and more precise as the number of samples t increased.

Increasing the number of samples has several favorable effects on the unmixing: the number of samples available to unmix the sources becomes sufficient regardless of the presence of outliers, and the clustered aspect of **AS** has a greater importance since there are more and more clustered samples in the term **AS** but the outliers are still in general position (generating randomly several outliers in a same direction is quite unlikely).

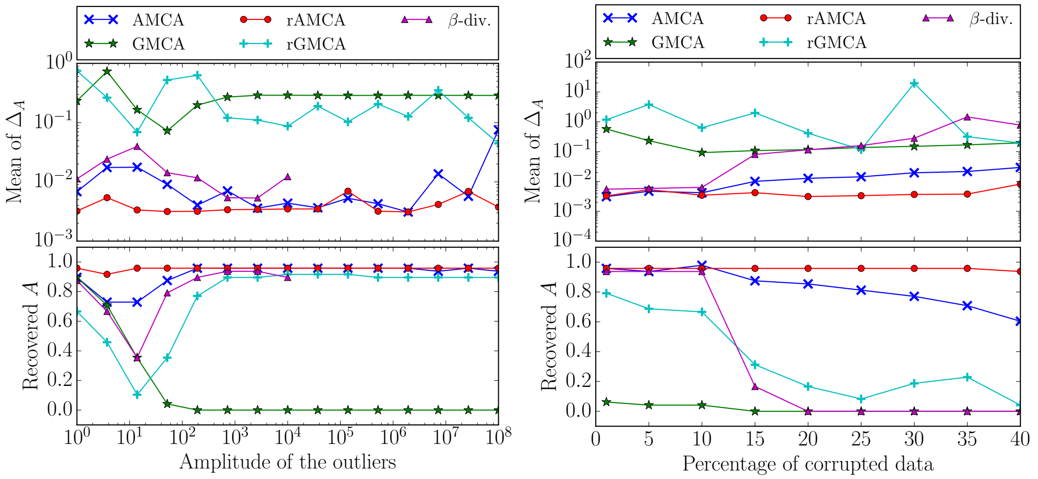
The results would have been different if the proportion of data samples in a given direction had been set constant from one value of t to another. For instance, if one resizes an multi/hyperspectral data cube, these proportions are kept constant, and for the largest image size, few but several outliers are in a same direction. There are some applications (most of the observations of physical processes) were the outliers are not strictly speaking in general position (such as in 4.6), but whose contributions are less structured/clustered than the one of the sources: the weighting scheme penalizes the less clustered solutions, and so, still returns $\mathbf{A}$. That is why AMCA and rAMCA requires less samples than the others methods to perform accurately fig.5.

In the following, the impact of two other parameters will be investigated: the percentage of corrupted data and their amplitudes. We will focus on the determined case which is more challenging. Since the low-rankness assumption makes no sense in the determined case, the algorithm PCP+GMCA will not be evaluated.

4.4. Influence of the amplitude of the outliers

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

The figure 6a 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, rAMCA tends to be the



(a) Influence of σ_O.

(b) Influence of ρ.

Figure 6: Influence of the amplitude and the activation rate of the outliers.

least impacted by the amplitude of the outliers, especially when their amplitude is of the order of the source's level or very large. When the amplitude of the outliers and the sources are close, the weighting schemes of AMCA and rGMCA are 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.

4.5. Influence of the percentage of corrupted data

In this section, the amplitudes of the outliers σ_O is fixed to 100. The figure 6b shows the behavior of the BSS algorithms when the percentage of corrupted columns ρ varies according to the values of the x-axis.

As illustrated in fig.6b, 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%). The rGMCA algorithm is rapidly impacted by an increasing number of corrupted data. On the other hand, the AMCA-based algorithms are less influenced by the percentage of outliers. The rAMCA algorithm provides a significantly better estimate of the mixing matrix when the number of outliers is larger than 10%.

4.6. Application to NMR spectra unmixing

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]. 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] with $t = 2048$ samples. 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, fig.7a. The set of corrupted data samples is fixed to 10 blocks of 20 consecutive columns. Their amplitudes are drawn according to a Chi-law with 1 degree of freedom, 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 $\frac{\|\mathbf{O}\|_2}{10} = \frac{\|\mathbf{AS}\|_2}{n}$, fig.7c. In the following experiments, the data

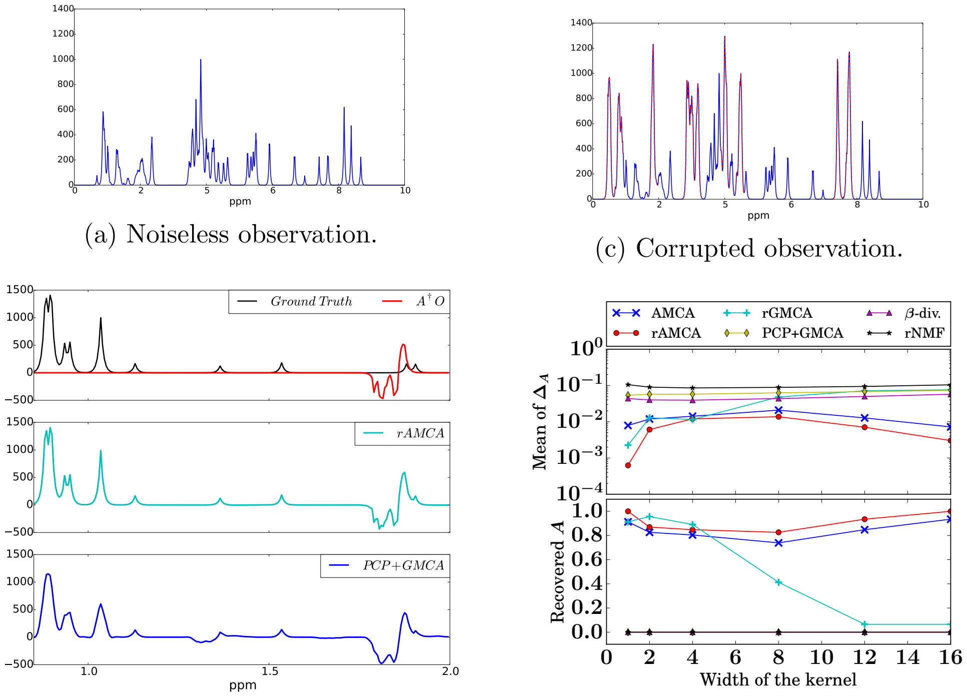
[1] http://sdbs.db.aist.go.jp

[2] http://www.cosmostat.org/software/gmcalab/

made of 10 mixtures computed with a positive mixing matrix their entries are drawn from a Chi-law with 1 degree of freedom and then the columns are normalized) and corrupted also by the presence of the centered Gaussian noise with $\sigma_N = 0.1$.

Given that all the variables are non-negative, we will also compare AMCA and rAMCA with rNMF [2], whose code is online. This method exploits the low-rankness of $\mathbf{AS}$, the non-negativity and the "sum-to-one"constraint (that is, the amplitudes of each sample of $\mathbf{S}$ sum to one) to differentiate between the low-rank subspace and the outliers. The "sum-to-one"constraint, which is not a valid assumption in this setting, is replaced by the constraint on the columns of $\mathbf{A}$, which are assumed to be normalized. We use the following inputs for rNMF: the ground truth $\mathbf{A}$, the projected sources $\left(\mathbf{A}^\dagger\mathbf{X}\right)_+$ and the non-negative part of the corresponding residue.

The resulting sources admit a sparser distribution in the wavelet domain. Subsequently, the data are transformed with the undecimated wavelet transform [28] prior to applying the BSS algorithms, except for rNMF. Let us notice that a same wavelet transform is used for the outliers and the sources because they have a similar morphology in the present setting. In the previous experiments, we evaluated the separation performances of 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. We propose to evaluate the behavior of 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



(a) Noiseless observation.

(c) Corrupted observation.

(b) Estimated source.

(d) Influence of the kernel width.

Figure 7: Top: illustration of one observation $\mathbf{X}_i$, without (left) and with outliers (right, corrupted entries are represented with the red dashed line). Bottom: estimated sources with rAMCA and PCP+GMCA for a width of the kernel of 6 (left) and right, performances of the different algorithms versus the width of the kernel (right).

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 7d displays the evolution of the mixing matrix criterion when the width of the convolution kernel varies. It is interesting to notice that the minimization of the β-divergence, PCP+GMCA, and the rNMF algorithms do not provide satisfactory separation results. This experience is particularly challenging for these methods since: the low-rank assumption is not valid, the "sum-to-one" constraint necessary to the separation between **AS** and **O** on **S** for rNMF has been removed, and the outliers are less and less sparse as the width of the kernel increases. 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 scenarios 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 [18]. Last, one of the sources estimated by rAMCA and PCP+GMCA are displayed in fig.7b. Contrary to PCP+GMCA, the source is correctly recovered by rAMCA outside the support of **O** because **A** is correctly estimated by rAMCA. However, the leakages from the outliers into the sources estimated by rAMCA are still important: they come from the coarse scale of the wavelet coefficients, which is not sparse and for which we cannot differentiate the two contributions. Taking into account the non-negativity of the signals would limit these leakages, but necessitates the use of proximal algorithms [29] if combined with sparsity in a transformed domain [26]. Nonetheless, the weighting scheme of rAMCA and AMCA is sufficient to obtain a robust estimation of **A**.

Software

Following the philosophy of reproducible research [30], a python implementation of the algorithms introduced in this article will be available at *http://www.cosmostat.org/GMCALab*.

5. Conclusion

In this article, we introduce a new algorithm for tackling BSS problems in the presence of outliers, which is a key problem in a large number of applications. The proposed rAMCA algorithm performs by estimating jointly the mixing matrix, the sources and the outliers. Inspired by the AMCA algorithm, it first provides a robust estimation of the sources and the mixing matrix. Additionally, it exploits the difference of structures of the outliers and the sources to provide a robust detection and estimation of the outliers based on their sparsity level in the source domain. Numerical experiments have been carried out on Monte-Carlo simulations with various experimental scenarios, which show that rAMCA yields a robust and reliable estimation of the mixing matrix. It provides the state-of-the-art separation results especially in the highly challenging determined case. Future work will exploit the difference of morphology between the sources and the outliers that is manifested in various imaging problems to further perform an accurately separation the two contributions.

Acknowledgment

This work is supported by the European Community through the grants PHySIS (contract no. 640174), DEDALE (contract no. 665044) and LENA

(ERC StG no. 678282) within the H2020 Framework Program.

[1] J. Bobin, F. Sureau, J.-L. Starck, A. Rassat, P. Paykari, Joint Planck and WMAP CMB map reconstruction, A&A 563 (A105).

[2] C. Fevotte, N. Dobigeon, Nonlinear Hyperspectral Unmixing With Robust Nonnegative Matrix Factorization, Image Processing, IEEE Transactions on 24 (12) (2015) 4810–4819. doi:10.1109/TIP.2015.2468177.

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

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

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

[6] Q. Li, H. Li, Z. Lu, Q. Lu, W. Li, Denoising of Hyperspectral Images Employing Two-Phase Matrix Decomposition, Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of 7 (9) (2014) 3742–3754. doi:10.1109/JSTARS.2014.2360409.

[7] P. G. Van Dokkum, Cosmic-ray rejection by laplacian edge detection, Publications of the Astronomical Society of the Pacific 113 (789) (2001) 1420.

[8] F. Sureau, J.-L. Starck, J. Bobin, P. Paykari, A. Rassat, Sparse point-source removal for full-sky cmb experiments: application to wmap 9-year data, Astronomy &amp; Astrophysics 566 (2014) A100.

[9] M. Mihoko, S. Eguchi, Robust blind source separation by beta divergence, Neural computation 14 (8) (2002) 1859–1886.

[10] E. J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, Journal of the ACM (JACM) 58 (3) (2011) 11.

[11] H. Zhang, W. He, L. Zhang, H. Shen, Q. Yuan, Hyperspectral Image Restoration Using Low-Rank Matrix Recovery, Geoscience and Remote Sensing, IEEE Transactions on 52 (8) (2014) 4729–4743. doi:10.1109/TGRS.2013.2284280.

[12] L. Zhang, Z. Chen, M. Zheng, X. He, Robust non-negative matrix factorization, Frontiers of Electrical and Electronic Engineering in China 6 (2) (2011) 192–200.

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

[14] C. Li, Y. Ma, X. Mei, C. Liu, J. Ma, Hyperspectral unmixing with robust collaborative sparse regression, Remote Sensing 8 (7) (2016) 588.

[15] A. Halimi, J. Bioucas-Dias, N. Dobigeon, G. S. Buller, S. McLaughlin, Fast hyperspectral unmixing in presence of nonlinearity or mismodelling effects, arXiv preprint arXiv:1607.05336.

[16] Y. Altmann, S. McLaughlin, A. Hero, Robust Linear Spectral Unmixing Using Anomaly Detection, IEEE Transactions on Computational Imaging 1 (2) (2015) 74–85. doi:10.1109/TCI.2015.2455411.

[17] C. Chenot, J. Bobin, J. Rapin, Robust Sparse Blind Source Separation, Signal Processing Letters, IEEE 22 (11) (2015) 2172–2176.

[18] J. Bobin, J. Rapin, A. Larue, J.-L. Starck, Sparsity and Adaptivity for the Blind Separation of Partially Correlated Sources, Signal Processing, IEEE Transactions on 63 (5) (2015) 1199–1213. doi:10.1109/TSP.2015.2391071.

[19] J.-L. Starck, F. Murtagh, J. M. Fadili, Sparse image and signal processing: wavelets, curvelets, morphological diversity, Cambridge University Press, 2010.

[20] M. Zibulevsky, B. Pearlmutter, Blind Source Separation by Sparse Decomposition in a Signal Dictionary, Neural Computation 13 (4) (2001) 863–882. doi:10.1162/089976601300014385.

[21] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, Journal of optimization theory and applications 109 (3) (2001) 475–494.

[22] M. Kowalski, Sparse regression using mixed norms, Applied and Computational Harmonic Analysis 27 (3) (2009) 303–324.

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

[24] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE transactions on pattern analysis and machine intelligence 31 (2) (2009) 210–227.

[25] N. Gadhok, 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.

[26] J. Rapin, J. Bobin, A. Larue, J.-L. Starck, NMF with Sparse Regularizations in Transformed Domains, SIAM Journal on Imaging Sciences 7 (4) (2014) 2020–2047.

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

[28] J.-L. Starck, J. Fadili, F. Murtagh, The undecimated wavelet decomposition and its reconstruction, IEEE Transactions on Image Processing 16 (2) (2007) 297–309.

[29] P. L. Combettes, V. R. Wajs, Signal recovery by proximal forward-backward splitting, Multiscale Modeling & Simulation 4 (4) (2005) 1168–1200.

[30] J. B. Buckheit, D. L. Donoho, Wavelets and Statistics, Springer New York, 1995, Ch. WaveLab and Reproducible Research, pp. 55–81.