HAL
open science

Blind Source Separation with Outliers in Transformed Domains

Cécile Chenot, Jerome Bobin

To cite this version:

Cécile Chenot, Jerome Bobin. Blind Source Separation with Outliers in Transformed Domains. SIAM Journal on Imaging Sciences, 2018, 11 (2), pp.1524-1559. 10.1137/17m1133919 . hal-03177865

HAL Id: hal-03177865

https://hal.science/hal-03177865

Submitted on 23 Mar 2021

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

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

Blind Source Separation with outliers in transformed domains*

Cécile Chenot † and Jérôme Bobin†

Abstract. Blind Source Separation (BSS) methods are well suited for the analysis of multichannel data. In many applications, the observations are corrupted by an additional structured noise, which hinders most of the standard BSS techniques. In this article, we propose a novel BSS method able to jointly unmix the sources and separate the source contribution from the structured noise or outliers. This separation builds upon the difference of morphology between the components of interest, often encountered in imaging problems, by exploiting a sparse modeling of the components in two different domains. Numerical experiments highlight the robustness and precision of the proposed method in a wide variety of settings, including the full-rank regime.

Key words. Blind Source Separation, Sparse Modeling, Robust Recovery, Morphological Diversity.

AMS subject classifications. 68U10

1. Introduction.

Blind Source Separation (BSS) is a powerful tool to extract the meaningful information of multichannel data, which are encountered in various domains such as biomedical engineering [39] or remote-sensing [3] to cite only a few. Notably, it has played a key role in the analysis of the multispectral observations of the ESA-Planck mission [7], [30] in astrophysics. Its instantaneous linear mixture model assumes that n sources $\{\mathbf{S}_i\}_{i=1..n}$ of t samples are mixed into $m \geq n$ observations $\{\mathbf{X}_j\}_{j=1..m}$. This model can be conveniently recast in the following matrix formulation:

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{m\times t}$ designates the linear observations, $\mathbf{A} \in \mathbb{R}^{m\times n}$ the unknown mixing matrix, $\mathbf{S} \in \mathbb{R}^{n\times t}$ the sources and $\mathbf{N} \in \mathbb{R}^{m\times t}$ a Gaussian noise term accounting for model imperfections. BSS aims at recovering both $\mathbf{A}$ and $\mathbf{S}$ from $\mathbf{X}$. This is an ill-posed problem as the number of solutions is infinite. Recovering the relevant sources and mixing matrix then requires additional prior information on the sources and/or the mixing matrix such as: the mutual independence of the sources in the ICA framework [17], the non-negativity of $\mathbf{S}$ and $\mathbf{A}$ for Non-negative Matrix Factorization (NMF) [35], or the compressibility of the sources in a given domain [57]. Further details on standard BSS can be found in [17] and references therein.

This model is too simple to represent accurately some complex processes. In the ESA-Planck mission for instance, the observations deviate from the above model 1 because of the presence of point-source emissions with unknown position and amplitude as well as the spectral variability of some components [30]. The spectral variability of some components is

*Submitted to the editors DATE.

Funding: This work is supported by the European Community through the grants PHySIS (contract no. 640174) and LENA (ERC StG no. 678282) within the H2020 Framework Program.

†CEA, Irfu, Service d'Astrophysique - SEDI, 91191 Gif-sur-Yvette Cedex, France (cecile.chenot@cea.fr,jerome.bobin@cea.fr).

This manuscript is for review purposes only.

also a major issue in hyperspectral imaging and has encountered a growing interest during the last years [52],[21]. More generally, deviations from the standard model 1 are encountered in numerous applications and encompass the presence of unexpected physical events [44], [47], instrumental artifacts [31] or non-linearity of the physical process [21]. These large errors will be designated in the following as *outliers*. In order to take into account these deviations in the data model, we propose to model the observations with the following expression:

$$\mathbf{X} = \mathbf{AS} + \mathbf{O} + \mathbf{N}, \tag{2}$$

where $\mathbf{O} \in \mathbb{R}^{m \times t}$ stands for the outliers.

Robust BSS in the literature. Most standard BSS methods lead to inaccurate or erroneous results in the presence of outliers [22]. This mandates the development of *robust* BSS methods, which should tackle both following tasks:

- Unmixing of the sources, *i.e.* estimating precisely the mixing matrix $\mathbf{A}$.
- Separating the source contribution $\mathbf{AS}$ from the outliers $\mathbf{O}$ so as to return non-corrupted sources.

Only few robust methods have been proposed in the literature. They can be classified into three different groups according to their strategies: replacement of the sensitive metrics in the cost-functions of optimization based-methods (*i.e.* only task i)), removal of the outliers prior to the unmixing (*i.e.* task ii) followed by task i)), and joint estimation of $\mathbf{O}$, $\mathbf{S}$ and $\mathbf{A}$ (*i.e.* tasks i) and ii) simultaneously).

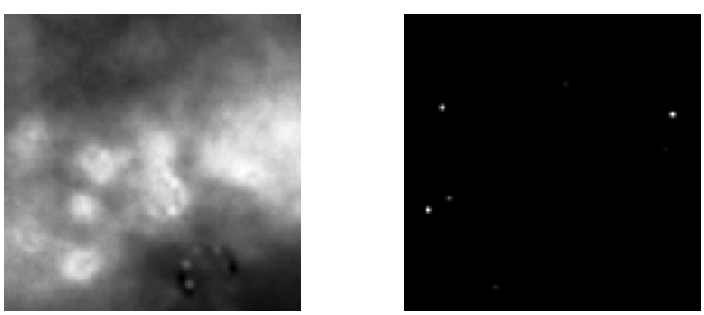
A robust unmixing of the sources without an explicit estimation of the outliers has been proposed in several works. This approach consists of replacing the most sensitive metrics of the cost functions of the usual optimization-based BSS methods. In the NMF framework for instance, the authors of [25] and [27] opt respectively for the ℓ_1 and the $\ell_{2,1}$ norms for the data fidelity term, instead of the common Frobenius norm which is sensitive to large errors. In the ICA framework, the authors of [32] promote the mutual independence of the sources by using the robust β-divergence in place of the Kullback-Leibler divergence [17]. The major drawback of this class of methods is that only the mixing matrix can be recovered precisely, while the sources are still contaminated by the outliers.

The second popular approach consists in: first, estimating and discarding the outliers from the observations, and then, performing the BSS on the denoised observations. In order to unmix the sources accurately in the second step, the estimation of the outliers should be very precise. However, this step is challenging and necessitates further assumptions (see Section 2). A popular strategy for discarding the outliers assumes that $m \gg n$ so that $\mathbf{AS}$ has low-rank. In [11], it has been proven that an exact separation between the outliers and $\mathbf{AS}$ is possible if the support of the outliers is, in addition, uniformly distributed. This approach has been in particular used in hyperspectral imaging for which the assumption on the low-rankness holds true [53]. The major drawback of this strategy is that it does not take into account explicitly the clustering aspect of $\mathbf{AS}$ and the assumption made on $\mathbf{S}$ for the following unmixing (*e.g.* independence or sparsely represented in a given dictionary) since $\mathbf{A}$ and $\mathbf{S}$ are not estimated explicitly. This can greatly hamper the unmixing by propagating the error made at the first

step.

The third class of methods estimates jointly $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{O}$. This allows to constrain all the components and limits the propagation of errors encountered with the previous two-steps methods. This strategy has been developed essentially with non-negativity [40], [54] and low-rank priors for hyperspectral unmixing [1], [21], [33]. In [14], we proposed a robust BSS method assuming that both the outliers and the sources are sparsely represented in a same domain. The proposed method estimates reliably the mixing matrix, but was however unable to separate precisely the sources and the outliers without additional assumption (see Section 2) in the full-rank setting.

Contributions. To the best of our knowledge, there is currently no BSS method able to estimate the mixing matrix, the sources and the outliers in a general framework *i.e.* without the low-rank assumption.
In this paper, we propose to exploit the difference of morphology/geometrical content between the outliers and the sources to separate precisely the two contributions [20]. This difference of morphology is often encountered in imaging problems: stripping lines due to malfunctions of captors have a different morphology than natural images in multi/hyperspectral imaging or point-source emissions fig.1b have a different geometry than the sought-after signals fig.1a in the ESA-Planck mission. By only assuming that the outliers and the sources have a different morphology, our new strategy coined tr-rGMCA (robust Generalized Morphological Component Analysis in transformed domains), preliminarily presented in [15], is able to separate precisely the sources and the outliers, in a wide variety of problems, including in the challenging determined case ($n = m$).



(a) Simulated synchrotron emission.

(b) Simulated point-source emissions.

Figure 1: Simulated components of the ESA-Planck mission: synchrotron's map (one row of $\mathbf{S}$) (a) and observation of the point-sources contamination at a given frequency (one row of $\mathbf{O}$) (b).

The structure of this article is the following: in Section II, we focus on the separation of the outliers from the sources contribution for which we explain why the morphological diversity

is a powerful assumption, in Section III, we introduce the tr-rGMCA problem, the associated algorithm and the strategies used for the automatic choice of the parameters, and last in Sections IV to VI, the results of numerical experiments on 1D Monte-Carlo simulations and 2D simulated astrophysics data are displayed for the comparison of tr-rGMCA with standard robust BSS methods.

Notations.

Matrix notations. Matrices are denoted by uppercase boldface letters. The ith row and jth column of a matrix $\mathbf{M}$ are designated respectively by $\mathbf{M}_i$ and $\mathbf{M}^j$, and its i,jth entry by $\mathbf{M}_{i,j}$. The Moore-Penrose inverse of $\mathbf{M}$ is noted $\mathbf{M}^\dagger$ and its transpose $\mathbf{M}^T$.
The notation $\tilde{\mathbf{M}}$ denotes the estimate of $\mathbf{M}$ and, $\tilde{\mathbf{M}}^{(k)}$ designates the estimate at the kth iteration of a loop.

Norms. Three 'entrywise'norms will be used: $\|\mathbf{M}\|_1 = \sum_{i,j} |\mathbf{M}_{i,j}|$, $\|\mathbf{M}\|_{2,1} = \sum_j \|\mathbf{M}^j\|_2$, and $\|\mathbf{M}\|_2$, the Frobenius norm of $\mathbf{M}$.

Operators. The operators $\odot, \circledast, \otimes$ designate the Hadamard product, the convolution, and the tensor product respectively.
The proximal operator of a real-valued, convex, proper and lower semicontinuous function $f : \mathbb{R}^{p\times r} \to \mathbb{R}$, $(p, r \in \mathbb{N})$, is noted prox_f, such that $\text{prox}_f : \mathbb{R}^{p\times r} \to \mathbb{R}^{p\times r}$, $\mathbf{X} \mapsto \text{argmin}_{\mathbf{Y}} \frac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_2^2 + f(\mathbf{Y})$.
In particular, the soft-thresholding operator of $\mathbf{M}$, with threshold λ, is denoted $\mathcal{S}_\lambda(\mathbf{M})$, where

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

Last, the operator **mad** designates the median absolute deviation.

2. Separation between the outliers and the sources.

Robust blind source separation can merely be split into two distinct problems: i) the robust estimation of the mixing matrix from the data without considering outliers removal and ii) the exact or accurate separation between the outliers $\mathbf{O}$ and the sources $\mathbf{AS}$. In this section, we discuss the properties needed to tackle these two problems.

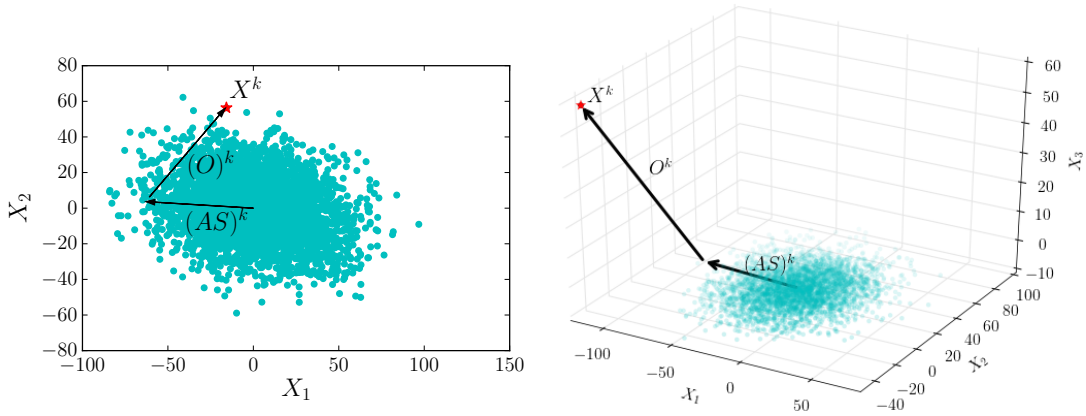
2.1. Spectral diversity.

A robust PCA perspective. In this paragraph, we first focus on the second problem: separating the contribution of the sources $\mathbf{AS}$ from the outliers $\mathbf{O}$. If one defines $\mathbf{L} = \mathbf{AS}$, the data can be described as

$$\mathbf{X} = \mathbf{L} + \mathbf{O}$$

Assuming that the outliers have a sparse distribution and are in general position (they do not cluster in a specific direction), and that $\mathbf{L}$ is low-rank (*i.e.* the number of observations is much larger than the number of sources $m \gg n$), this separation problem refers to robust PCA (rPCA - [11], [13]). To illustrate this particular setting, we display in fig.3, the scatter plot of the observations in a determined setting fig.2b and an over-determined setting fig.2b. Intuitively, the fact that corrupted samples do not lie in the span of $\mathbf{L}$ facilitates their detection as illustrated in fig.2b. This is however not a sufficient condition for the identifiability of the

two components.



(a) Determined case. (b) Over-determined case.

Figure 2: On the left, 2 sources are mixed into 2 corrupted observations. On the right, 2 sources are mixed into 3 observations. For both, the red star symbolizes the corrupted sample, at the kth column and the arrows symbolize the two contributions to this sample $\mathbf{O}^k$ and $(\mathbf{AS})^k$.

It was shown in [11] and [13] that the identifiability of the components can be proved if i) the entries of $\mathbf{O}$ are sparse, in general position and independently distributed ($\mathbf{O}$ is row and column sparse) and ii) the component $\mathbf{L}$ lies in a low-dimensional subspace, with broadly distributed entries ($\mathbf{L}$ is not row or column sparse).
The PCP algorithm, designed to perform this separation [11], estimates both components by solving:

$$\min_{\mathbf{L},\mathbf{O}:\mathbf{X}=\mathbf{L}+\mathbf{O}} \|\mathbf{L}\|_* + \lambda \|\mathbf{O}\|_1 . \tag{3}$$

where $\|\mathbf{L}\|_*$ stands for the nuclear norm of $\mathbf{L}$ (*i.e.* the sum of its singular values).

However, in the framework of rPCA, the exact or accurate [56] separation between the sources and the outliers is guaranteed as long as the entries of $\mathbf{O}$ are independently distributed, which excludes column-sparse outliers. For this case, PCP has been extended in [49]. The outliers pursuit (OP) algorithm minimizes:

$$\min_{\mathbf{L},\mathbf{O}:\mathbf{X}=\mathbf{L}+\mathbf{O}} \|\mathbf{L}\|_* + \lambda \|\mathbf{O}\|_{2,1} . \tag{4}$$

Interestingly, it has been shown that the OP algorithm allows retrieving the support of the outliers and the column span of $\mathbf{A}$. However, the separation of $\mathbf{L}$ and $\mathbf{O}$ is not guaranteed since the contribution of the outliers that lies in the span of $\mathbf{A}$ cannot be recovered exactly.

In the framework of robust BSS. Both the rPCA and OP algorithms strongly rely on the low-rankness of the source contribution in the data. This assumption makes perfect sense in applications such as hyperspectral imaging [21, 45], where few sources (typically $n < 10$) have to be estimated from a large number of observations (*i.e.* $m \sim 10^2$). However, the so-called hyperspectral unmixing methods take advantage of additional constraints to improve the separation between $\mathbf{AS}$ and $\mathbf{O}$ [3], [26]: i) the non-negativity of the mixing matrix $\mathbf{A}$ and the sources $\mathbf{S}$ and ii) the sources samples (*i.e* the columns of the sources matrix) are assumed to lie on the ℓ_1 simplex.

Unfortunately, neither the low-rankness nor the non-negativity assumptions are valid in a broad range of applications such as the Planck data. For that purpose, we introduced in [14] a robust BSS algorithm coined rAMCA that jointly estimates $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{O}$. The rAMCA algorithm builds upon the sparse modeling of the sources and the outliers in the same dictionary. If the rAMCA algorithm has been shown to outperform the state-of-the-art robust BSS methods including in the determined setting, it only provides a robust estimation of the mixing matrix and fails at accurately separating the sources and the outliers. Indeed, whether the low-rankness of the sources holds or not, column sparse outliers are not identifiable, which makes the sources/outliers separation impossible without additional assumptions.

2.2. Combining spectral and morphological diversity. In this section, we introduce an additional property that helps differentiating between the sources and the outliers: morphological diversity. While spectral diversity refers to the relative distributions of the sources and the outliers in the column-space, morphological diversity deals with their relative distribution in the row-space. Morphological diversity has first been exploited in the monochannel case to separate multiple images that share different geometrical structures. In that context, it has been quite successful at separating contour and texture parts in images. This concept is at the origin of the MCA algorithm (Morphological Component Analysis - [20, 6]).
In a large number of applications, the sources to be retrieved and the outliers share different morphologies, such as in Planck data fig.1. In this case, spurious points sources are the perfect example of column sparse outliers. These components are local singularities that are morphologically distinct from more diffuse astrophysical components. Therefore, building upon the concept of morphological diversity, we hereafter propose to reformulate robust BSS as special case of multichannel MCA problem. In the remaining of this paper, we will make use of the following assumptions:

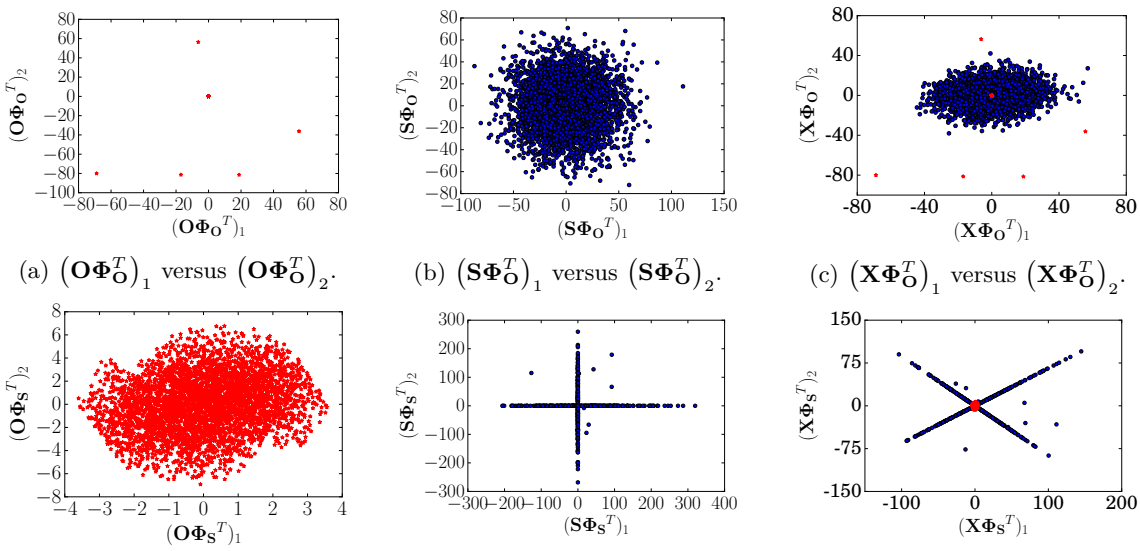
- **Morphological diversity between the sources and the outliers:** We assume that the sources are sparsely represented in the transformed domain or dictionary $\mathbf{\Phi_S}$ and that the outliers have a sparse representation in $\mathbf{\Phi_O}$:

$$\mathbf{O}_j = \boldsymbol{\alpha}_{\mathbf{O}_j}\mathbf{\Phi_O},\ \forall j \in \{1..m\} \quad \text{and} \quad \mathbf{S}_i = \boldsymbol{\alpha}_{\mathbf{S}_i}\mathbf{\Phi_S},\ \forall i \in \{1..n\}\,,$$

  where $\left\{\boldsymbol{\alpha}_{\mathbf{O}_j}\right\}_{j=1..m}$ and $\{\boldsymbol{\alpha}_{\mathbf{S}_i}\}_{i=1..n}$ are composed of few significant samples. These dictionaries should be chosen according to the main structural characteristics of the

components to assure that the expansion coefficients are sparse, *e.g.* wavelets for piecewise smooth signals, curvelets for curves like cartoons or DCT for oscillating textures [43]. A toy example is provided in fig.3. It highlights the benefits of exploiting the morphological diversity: in $\mathbf{\Phi_S}$, the outlier contribution is broadly distributed with a very small amplitude fig.3d,3f, whereas in $\mathbf{\Phi_O}$, they can be easily detected fig.3a,3c (and reciprocally for the sources samples).

- **Sparse modeling of the outliers:** We also consider that the sparse representations of the outliers corrupt entirely some columns and are broadly distributed in all the directions. For this purpose, we will assume that $\mathbf{O\Phi_O^T}$ is column sparse such as in fig.3. For instance, in the applications for which the outliers are sparse in the domain of observation, it amounts supposing that most of the sensors record the spurious outliers at a same instant/position: that is the case of the point source emissions in astrophysics fig.1b. We point out that assuming that the outliers are column and row sparse in $\mathbf{\Phi_O}$ only requires minor changes, which will be indicated in the following.



(a) $\left(\mathbf{O\Phi_O^T}\right)_1$ versus $\left(\mathbf{O\Phi_O^T}\right)_2$. (b) $\left(\mathbf{S\Phi_O^T}\right)_1$ versus $\left(\mathbf{S\Phi_O^T}\right)_2$. (c) $\left(\mathbf{X\Phi_O^T}\right)_1$ versus $\left(\mathbf{X\Phi_O^T}\right)_2$.

(d) $\left(\mathbf{O\Phi_S^T}\right)_1$ versus $\left(\mathbf{O\Phi_S^T}\right)_2$. (e) $\left(\mathbf{S\Phi_S^T}\right)_1$ versus $\left(\mathbf{S\Phi_S^T}\right)_2$. (f) $\left(\mathbf{X\Phi_S^T}\right)_1$ versus $\left(\mathbf{X\Phi_S^T}\right)_2$.

Figure 3: Two sources sparse in DCT are mixed into three observations, corrupted with sparse outliers. (a): scatter plot of the two first rows of $\mathbf{O}$ in $\mathbf{\Phi_O}$, (d): scatter plot the same rows of $\mathbf{O}$ in $\mathbf{\Phi_S}$, (b): scatter plot of the first two sources in $\mathbf{\Phi_O}$, (e): scatter plot of the same sources in $\mathbf{\Phi_S}$, (c): scatter plot of the two first corrupted observations in $\mathbf{\Phi_O}$ and last, (f): scatter plot of the same observations in $\mathbf{\Phi_S}$. The source contribution is represented with the blue dots and the outliers with the red stars.

2.2.1. Robust (non-blind) source separation as a sparse decomposition problem.

A special case of sparse decomposition in an overcomplete dictionary. Following standard sparse BSS approaches [57, 5], the sources are assumed to be sparsely distributed in a signal

representation or dictionary $\mathbf{\Phi_S}$ such that $\mathbf{S} = \boldsymbol{\alpha}_\mathbf{S}\mathbf{\Phi_S}$. The sources' contribution $\mathbf{L} = \mathbf{AS}$ to the data $\mathbf{X}$ is therefore sparsely represented in the multichannel dictionary: $\mathbf{A} \otimes \mathbf{\Phi_S}$, whose atoms are composed of tensor products between the columns of $\mathbf{A}$ and the atoms of $\mathbf{\Phi_S}$.

Similarly, the rows of the outlier matrix $\mathbf{O}$ are assumed to be sparse in some dictionary $\mathbf{\Phi_O}$ so that $\mathbf{O} = \boldsymbol{\alpha}_\mathbf{O}\mathbf{\Phi_O}$, where the coefficients $\boldsymbol{\alpha}_\mathbf{O}$ are column sparse. Let $\mathbf{O}_D$ be the submatrix made of the normalized non-zero columns of $\boldsymbol{\alpha}_\mathbf{O}$, built so that the kth non-zero column of $\boldsymbol{\alpha}_\mathbf{O}$ at the position t equals $\mathbf{O}_D^k = \frac{\boldsymbol{\alpha}_\mathbf{O}{}^t}{\|\boldsymbol{\alpha}_\mathbf{O}{}^t\|_2}$. We then denote $\boldsymbol{\alpha}_{\mathbf{O}\prime}$ the expansion coefficients of $\mathbf{O}$ in $\mathbf{O}_D \otimes \mathbf{\Phi_O}$, such that $\mathbf{O}_D\boldsymbol{\alpha}_{\mathbf{O}\prime} = \boldsymbol{\alpha}_\mathbf{O}$. The matrix $\boldsymbol{\alpha}_{\mathbf{O}\prime}$ is then column and row sparse and $\|\boldsymbol{\alpha}_{\mathbf{O}\prime}\|_1 = \|\boldsymbol{\alpha}_\mathbf{O}\|_{2,1}$. With this parameterization, the outliers are sparsely represented in the multichannel dictionary $\mathbf{O}_D \otimes \mathbf{\Phi_O}$.

The observations are consequently sparsely represented in the multichannel dictionary $\mathbf{D} = [\mathbf{A} \otimes \mathbf{\Phi_S}, \mathbf{O}_D \otimes \mathbf{\Phi_O}]$:

$$\mathbf{X} = \begin{bmatrix}\mathbf{A} & \mathbf{O}_D\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha}_\mathbf{S} & 0 \\ 0 & \boldsymbol{\alpha}_{\mathbf{O}\prime}\end{bmatrix} \begin{bmatrix}\mathbf{\Phi_S} \\ \mathbf{\Phi_O}\end{bmatrix}.$$

Assuming that $\mathbf{A}$ and $\mathbf{O}_D$ are known, estimating the sources $\mathbf{S}$ and the outliers $\mathbf{O}$ from $\mathbf{X}$ boils down to tackling a sparse decomposition problem in the overcomplete multichannel dictionary $\mathbf{D}$. In the very large literature devoted to sparse decompositions in overcomplete dictionaries (see [9] for a review), different approaches have been proposed to investigate the identifiability and recovery of sparse decompositions. In the next, we make use of the so-called mutual coherence of the dictionary to provide a deeper insight into the proposed robust component separation.
Assuming that the components are K-sparse in $\mathbf{D}$ with $K = \|\boldsymbol{\alpha}_\mathbf{S}\|_0 + \|\boldsymbol{\alpha}_{\mathbf{O}\prime}\|_0$, a sufficient condition for the identifiability of $\boldsymbol{\alpha}_\mathbf{S}$ and $\boldsymbol{\alpha}_{\mathbf{O}\prime}$ [19] is given by:

$$K < \frac{1}{2}\left(1 + \frac{1}{\mu_\mathbf{D}}\right),$$

where $\mu_\mathbf{D}$ designates the so-called mutual coherence of the dictionary $\mathbf{D}$. The mutual coherence of $\mathbf{D}$ is defined as $\mu_\mathbf{D} = \max_{i,j}|\langle \mathbf{d}_i, \mathbf{d}_j\rangle|$ where $\mathbf{d}_i$ stands for an atom of the multichannel dictionary $\mathbf{D}$ (*i.e.* multichannel atoms are composed of tensor products of atoms from the spectral dictionaries and morphological dictionaries). Furthermore, the same condition also guarantees that $\boldsymbol{\alpha}_\mathbf{S}$ and $\boldsymbol{\alpha}_{\mathbf{O}\prime}$ can be recovered by solving the following basis pursuit problem [19]:

$$\underset{\boldsymbol{\alpha}_{\mathbf{O}\prime},\boldsymbol{\alpha}_\mathbf{S}}{\operatorname{argmin}} \|\boldsymbol{\alpha}_{\mathbf{O}\prime}\|_1 + \|\boldsymbol{\alpha}_\mathbf{S}\|_1 \text{ s.t. } \mathbf{X} = \mathbf{A}\boldsymbol{\alpha}_\mathbf{S}\mathbf{\Phi_S} + \mathbf{O}_D\boldsymbol{\alpha}_{\mathbf{O}\prime}\mathbf{\Phi_O}. \tag{5}$$

The term of interest in the above recovery condition is the mutual coherence $\mu_\mathbf{D}$, which is equal, in this specific case, to:

$$\max\left(\max_{(i,p)\neq(j,q)}|\langle \mathbf{A}^i, \mathbf{A}^j\rangle||\langle \mathbf{\Phi}_\mathbf{S}^p, \mathbf{\Phi}_\mathbf{S}^q\rangle|, \max_{(m,u)\neq(n,v)}|\langle \mathbf{O}_D^m, \mathbf{O}_D^n\rangle||\langle \mathbf{\Phi}_\mathbf{O}^u, \mathbf{\Phi}_\mathbf{O}^v\rangle|, \max_{(l,c),(k,d)}|\langle \mathbf{A}^l, \mathbf{O}_D^k\rangle||\langle \mathbf{\Phi}_\mathbf{S}^c, \mathbf{\Phi}_\mathbf{O}^d\rangle|\right), \tag{6}$$

where the columns of $\mathbf{A}$, $\mathbf{\Phi_O}$ and $\mathbf{\Phi_S}$ are normalized to have unit ℓ_2 norm. In this expression, the cross-terms $\max_{(l,c),(k,d)} |\langle \mathbf{A}^l, \mathbf{O}_D^k \rangle| |\langle \mathbf{\Phi}_\mathbf{S}^c, \mathbf{\Phi}_\mathbf{O}^d \rangle|$ are the most relevant to discriminate the outliers and the sources' contribution and provide a different way to re-interpret robust (non-blind) source separation:

- **Spectral diversity or rPCA regime:** In case the outliers and sources share a same morphology, (see [29] for more precise recovery guarantees), only the cross-term between the mixing matrix and the outlier columns $\max_{(l,k)} |\langle \mathbf{A}^l, \mathbf{O}_D^k \rangle|$ is relevant for the separation. In the framework of rPCA, whenever the source contribution $\mathbf{AS}$ has low rank, $\max_{(l,k)} |\langle \mathbf{A}^l, \mathbf{O}_D^k \rangle|$ vanishes when $\mathbf{O}_D$ lies in the subspace that is orthogonal to the span of $\mathbf{A}$, which naturally ensures the identifiability of both the sources and the outliers. In the general case, assuming that $\mathbf{O}$ has independently and sparsely distributed entries and that $\mathbf{A}$ is broadly distributed such as in the setting of rPCA, leads to spectral dictionaries $\mathbf{A}$ and $\mathbf{O}_D$ with low coherence. This is precisely in this regime that rPCA can ensure the identifiability of the components.

- **Morphological diversity or MCA regime:** When the low-rankness of the observations is not a valid assumption or when the span of $\mathbf{A}$ and $\mathbf{O}_D$ are not incoherent, such as in the determined case, only the morphological diversity can help identifying the components. In that case, the dictionaries $\mathbf{\Phi_O}$ and $\mathbf{\Phi_S}$ are assumed to be incoherent, which makes $\max_{(c,d)} |\langle \mathbf{\Phi}_\mathbf{S}^c, \mathbf{\Phi}_\mathbf{O}^d \rangle|$ the relevant term for the separation. The is precisely in this regime that the MCA can ensure the separation between components that can only be identified thanks to their difference of morphologies. In this case only, robust component separation can be solved in the determined case.

- **Morpho/Spectral diversity:** In the general case, both the spectral and morphological dictionaries are incoherent, the relevant coherence term is the product $\max_{(l,c),(k,d)} |\langle \mathbf{A}^l, \mathbf{O}_D^k \rangle| |\langle \mathbf{\Phi}_\mathbf{S}^c, \mathbf{\Phi}_\mathbf{O}^d \rangle|$. In this regime, robust component separation benefits from incoherence of both the morphological and spectral dictionaries: $\max_{(l,c),(k,d)} |\langle \mathbf{A}^l, \mathbf{O}_D^k \rangle| |\langle \mathbf{\Phi}_\mathbf{S}^c, \mathbf{\Phi}_\mathbf{O}^d \rangle| \leq \min(\max_{(l,k)} |\langle \mathbf{A}^l, \mathbf{O}_D^k \rangle|, \max_{(c,d)} |\langle \mathbf{\Phi}_\mathbf{S}^c, \mathbf{\Phi}_\mathbf{O}^d \rangle|)$. This is expected to greatly improve the accuracy of the separation. For instance, in this regime, column-sparse outliers can be identified while methods that only make use of the spectral diversity like Outliers Pursuit [49] can only ensure the identification of the support of the outliers and not their amplitude.

In the framework of robust *blind* source separation, the spectral dictionary $\begin{bmatrix} \mathbf{A} & \mathbf{O_D} \end{bmatrix}$ is not known and has also to be learned. For this purpose, we describe in the next section a novel algorithm coined tr-rGMCA that makes use of both spectral and morphological diversity to estimate jointly $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{O}$ given the two dictionaries $\mathbf{\Phi_S}$ and $\mathbf{\Phi_O}$ so as to build upon the spectral and the morphological diversities between the components. Based on whether they rely on spectral or morphological diversity, currently available blind separation strategies are summarized in table.1.

Estimation	Diversity	Regime	Methods	Advantages	Weaknesses
AS and **O**	Morphological	$m \geq n$	MCA [20]	No assumption on the collinearity of **O** and **A**.	**AS** should be sparse in $\mathbf{\Phi_S}$. The spectral structure may not be preserved.
	Spectral	$m \gg n$	PCP [11], or refinements such as [55], [34]	Proven separability.	$\mathbf{O\Phi_O^T}$ column and row sparse.
			OP [49]	$\mathbf{O\Phi_O^T}$ column sparse.	No identifiability of **O**.
A, **S** and **O**	Spectral	$m \gg n$	rNMF [21]	Well adapted for hyperspectral unmixing.	Non-negativity, sources samples in the simplex and presence of almost pure-pixels.
		$m \geq n$	rAMCA [14]	Estimation of **A**	No identifiability of **O**.
	Morphological & spectral	$m \geq n$	**tr-rGMCA**	Identifiability in all regimes	

Table 1: Strategies able to separate **AS** and **O**.

3. Robust GMCA in transformed domains. In this section, we introduce the tr-rGMCA (rGMCA in transformed domains) algorithm that builds upon both spectral and morphological diversities to estimate simultaneously **A**, **S** and **O**. The tr-rGMCA algorithm performs the separation by minimizing a cost function whose elements are based on the following assumptions:

- Data fidelity term: The data are assumed to be described by the linear mixture model $\mathbf{X} = \mathbf{AS} + \mathbf{O} + \mathbf{N}$. The squared Frobenius norm $\|\mathbf{X} - \mathbf{AS} - \mathbf{O}\|_2^2$ is used as fidelity term to measure a discrepancy between the data and the model. This distance is well suited to account for the additive Gaussian noise **N** that usually contaminates the data.
- Penalty term for the sources: In the spirit of sparse BSS [57, 5], the sources are assumed to be sparsely represented in some dictionary $\mathbf{\Phi_S}$. The compressibility of **S** in $\mathbf{\Phi_S}$ is enforced with a weighted ℓ_1 norm of the expansion coefficients of **S**: $\left\|\Lambda \odot \mathbf{S\Phi_S^T}\right\|_1$. The weighting matrix $\Lambda \in \mathbb{R}^{n \times t_{\mathbf{\Phi_S}}}$ includes both the regularization parameters and the weights defined in standard re-weighting ℓ_1 penalization [12].
- Penalty term for the outliers: The outliers are assumed to be column-sparse in $\mathbf{\Phi_O}$. This structure is enforced in the cost function using the composite $\ell_{2,1}$ norm [21, 28]: $\left\|\Upsilon \odot \mathbf{O\Phi_O^T}\right\|_{2,1}$. Again the matrix $\Upsilon \in \mathbb{R}^{1 \times t_{\mathbf{\Phi_O}}}$, taille a relier aux dimensions des

dicos where contains the regularization parameters as well as weights in the sense of re-weighting $\ell_{2,1}$. The morphological diversity assumption implies that $\mathbf{\Phi_O}$ and $\mathbf{\Phi_S}$ are somehow incoherent.

- Scaling indeterminacy: In order to control the scaling indeterminacy between $\mathbf{A}$ and $\mathbf{S}$, the columns of $\mathbf{A}$ have an energy bounded by 1. The columns of $\mathbf{A}$ are constrained to lie in the ℓ_2 ball with unit radius: $\chi_{\mathbf{Y}:\|\mathbf{Y}^k\|_2\leq 1,\forall k}(\mathbf{A})$.

Therefore, the algorithm tr-rGMCA estimates jointly $\mathbf{A}$, $\mathbf{O}$ and $\mathbf{S}$ by minimizing the following cost function:

$$\underset{\mathbf{A},\mathbf{S},\mathbf{O}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{X}-\mathbf{AS}-\mathbf{O}\|_2^2+\left\|\Lambda\odot\mathbf{S}\mathbf{\Phi}_\mathbf{S}^T\right\|_1+\left\|\Upsilon\odot\mathbf{O}\mathbf{\Phi}_\mathbf{O}^T\right\|_{2,1}+\chi_{\mathbf{Y}:\|\mathbf{Y}^k\|_2\leq 1,\forall k}(\mathbf{A})\,. \tag{7}$$

The resulting cost function is a multi-convex *non-smooth* optimization problem: it is globally non-convex but subproblems with all variables fixed except one are convex. Hence, it is customary to optimize this type of cost function by iteratively and alternately minimizing it for each variable $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{O}$ assuming the others are fixed (namely the Block Coordinate optimization strategy, see [41] for a review). In particular, this is used by two standard strategies: the Block Coordinate Descent method (BCD - [46]), and Proximal Alternating Linearized Minimization (PALM - [8]).

3.1. Block Coordinate Minimization.

Updating each block $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{O}$ alternately at each iteration can be carried out in different ways. In the BCD setting, each block is updated by exactly minimizing Problem 7 assuming all the other blocks are fixed to their current values:

$$\mathbf{P}_A:\qquad \underset{\mathbf{A}}{\text{minimize}}\frac{1}{2}\|\mathbf{X}-\mathbf{AS}-\mathbf{O}\|_2^2+\chi_{\mathbf{Y}:\|\mathbf{Y}^k\|_2\leq 1,\forall k}(\mathbf{A})\,. \tag{8}$$

$$\mathbf{P}_S:\qquad \underset{\mathbf{S}}{\text{minimize}}\frac{1}{2}\|\mathbf{X}-\mathbf{AS}-\mathbf{O}\|_2^2+\left\|\Lambda\odot\mathbf{S}\mathbf{\Phi}_\mathbf{S}^T\right\|_1\,. \tag{9}$$

$$\mathbf{P}_O:\qquad \underset{\mathbf{O}}{\text{minimize}}\frac{1}{2}\|\mathbf{X}-\mathbf{AS}-\mathbf{O}\|_2^2+\left\|\Upsilon\odot\mathbf{O}\mathbf{\Phi}_\mathbf{O}^T\right\|_{2,1}\,. \tag{10}$$

These three problems can be written as $\text{argmin}_\mathbf{Y}\, f_Y(\mathbf{Y})+g_Y(\mathbf{Y})$, where $f_Y(.)$ is related to the differentiable data-fidelity term (whose gradient noted ∇f_Y is L_Y-Lipschitz) and $g_Y(.)$ is the proximable regularization associated with the component $\mathbf{Y}$ App. A. In general, they do not admit a closed-form solution and therefore require resorting to iterative minimization procedures such the Proximal Forward-Backward Splitting algorithm (FB) [16], [36]. In that case, BCD yields a computationally intensive minimization strategy. In the sequel, we therefore opted for the prox-linear approach, which is at the origin of the PALM algorithm [8]. In this framework, the PALM strategy updates each variable using a single proximal gradient step (it minimizes exactly the proximal linearization of each subproblem $\mathbf{P}_A$, $\mathbf{P}_S$ and $\mathbf{P}_O$, [50]). Whether it is based on BCD or PALM, it is possible to design a minimizer that provably converges to a local stationary point of Problem 7. In this context, either the BCD or the PALM algorithm can be chosen. However, the PALM procedure seems to generally converge faster: this can be understood with the fact that the components are updated with only one

proximal descent step, and not until convergence of each variable independently as done by BCD. For that reason, we opted for a prox-linear or PALM-based approach to design the algorithm.

3.2. A prox-linear implementation. In the framework of PALM, each component is updated with one proximal gradient step eq.11 at the kth iteration:

$$\tilde{\mathbf{Y}}^{(k)} \leftarrow \text{prox}_{\frac{1}{L_Y} g_Y}(\tilde{\mathbf{Y}}^{(k-1)} - \frac{1}{L_Y}\nabla f_Y((\tilde{\mathbf{Y}}^{(k-1)})). \tag{11}$$

From this generic update, the three steps that compose the tr-rGMCA algorithms are described as follows:

- **Update of the sources**. Assuming $\mathbf{\Phi_S}$ is orthonormal, the proximal operator of the function $\mathbf{S} \mapsto \left\|\Lambda \odot \mathbf{S}\mathbf{\Phi}_\mathbf{S}^T\right\|_1$ is exactly $\mathbf{S} \mapsto \mathcal{S}_\Lambda\Big((\mathbf{S})\,\mathbf{\Phi}_\mathbf{S}^T\Big)\,\mathbf{\Phi_S}$. Therefore, at iteration k of the PALM procedure, the update of $\tilde{\mathbf{S}}^{(k)}$ is given by:

$$\tilde{\mathbf{S}}^{(k+1)} \leftarrow \mathcal{S}_{\frac{\Lambda}{L_S}}\left(\left(\tilde{\mathbf{S}}^{(k)} + \frac{1}{L_S}\tilde{\mathbf{A}}^{(k)T}\left(\mathbf{X} - \tilde{\mathbf{A}}^{(k)}\tilde{\mathbf{S}}^{(k)} - \tilde{\mathbf{O}}^{(k)}\right)\right)\mathbf{\Phi}_\mathbf{S}^T\right)\mathbf{\Phi_S}, \tag{12}$$

  where the step size L_S is chosen to be equal to the Lipschitz constant of the gradient, *i.e.* the maximal eigenvalue of $\tilde{\mathbf{A}}^{(k)T}\tilde{\mathbf{A}}^{(k)}$.
  When $\mathbf{\Phi_S}$ is not orthonormal, the proximal operator of the function $\mathbf{S} \mapsto \left\|\Lambda \odot \mathbf{S}\mathbf{\Phi}_\mathbf{S}^T\right\|_1$ does not admit a closed form. However, in the next experiments, the dictionaries used of $\mathbf{\Phi_S}$ will be tight frames (*e.g.* undecimated wavelets) whose Gram matrix is close to the identity matrix. In that specific case, the update (12) provides a good approximation for the proximal operator.

- **Update of the outliers**. Assuming $\mathbf{\Phi_O}$ is orthonormal, the update of the outliers at the kth iteration of the PALM procedure is given by:

$$\tilde{\mathbf{O}}^{(k+1)} \leftarrow \tilde{\boldsymbol{\alpha}_\mathbf{O}}\mathbf{\Phi_O} \text{ where, } \forall j = 1..t \text{ and } \forall i = 1..n: \tag{13}$$

$$\tilde{\boldsymbol{\alpha}_\mathbf{O}}_i^j \leftarrow \left(\left(\left(\mathbf{X} - \tilde{\mathbf{A}}^{(k)}\tilde{\mathbf{S}}^{(k)}\right)\mathbf{\Phi}_\mathbf{O}^T\right)_i^j \times \max\left(0, 1 - \frac{\Upsilon_i^j}{\left\|\left(\left(\mathbf{X} - \tilde{\mathbf{A}}^{(k)}\tilde{\mathbf{S}}^{(k)}\right)\mathbf{\Phi}_\mathbf{O}^T\right)^j\right\|_2}\right)\right).$$

  In contrast to $\mathbf{S}$, the proximal gradient step eq.13 exactly solves $\mathbf{P}_O$. If $\mathbf{\Phi_O}$ is not orthogonal, but has a Gram matrix close to the identity, this update provides also a good approximation of the proximal gradient step. Besides, in this paper, we assume that the outliers corrupt entirely few columns of the observations in their associated transformed domain. However, it would be straightforward to account for row and column sparse outliers in $\mathbf{\Phi_O}$ by replacing the $\ell_{2,1}$ norm with the ℓ_1 norm. In this case, (13) is simply replaced by: $\tilde{\mathbf{O}} \leftarrow \mathcal{S}_\Upsilon\Big((\mathbf{X} - \mathbf{AS})\,\mathbf{\Phi}_\mathbf{O}^T\Big)\,\mathbf{\Phi_O}$.

- **Update of the mixing matrix**. The proximal gradient step for $\mathbf{A}$ is two step: ((a) corresponds to the gradient step, and (b) to the proximal operator of the characteristic function):

$$(a) \quad \tilde{\mathbf{A}}^{(k+1)} \leftarrow \tilde{\mathbf{A}}^{(k)} + \frac{1}{L_A}(\mathbf{X} - \tilde{\mathbf{A}}^{(k)}\tilde{\mathbf{S}}^{(k)} - \tilde{\mathbf{O}}^{(k)})\tilde{\mathbf{S}}^{(k)T},$$

$$(b) \quad \tilde{\mathbf{A}}^{i(k+1)} \leftarrow \frac{\tilde{\mathbf{A}}^{i(k+1)}}{\max\left(1, \left\|\tilde{\mathbf{A}}^{i(k+1)}\right\|_2\right)}, \forall i = 1..n,$$

where L_A is chosen to be equal to the Lipschitz constant of the gradient, *i.e.* the maximal eigenvalue of $\tilde{\mathbf{S}}^{(k)}\tilde{\mathbf{S}}^{(k)T}$.

Algorithm. The prox-linear minimization of (7) can be found in Alg.1.

Algorithm 1 PALM

procedure PALM$(\mathbf{X}, \tilde{\mathbf{S}}, \mathbf{\Phi_S}, \tilde{\mathbf{O}}, \mathbf{\Phi_O}, \tilde{\mathbf{A}}, \Lambda, \Upsilon)$
  Set $\tilde{\mathbf{S}}^{(0)} \leftarrow \tilde{\mathbf{S}}$, $\tilde{\mathbf{O}}^0 \leftarrow \tilde{\mathbf{O}}$ and $\tilde{\mathbf{A}}^0 \leftarrow \tilde{\mathbf{A}}$
  while $p < P$ **do**
    Compute L_A
    $\tilde{\mathbf{A}}^{(p+1)} \leftarrow \tilde{\mathbf{A}}^{(p)} + \frac{1}{L_A}(\mathbf{X} - \tilde{\mathbf{A}}^{(p)}\tilde{\mathbf{S}}^{(p)} - \tilde{\mathbf{O}}^{(p)})\tilde{\mathbf{S}}^{(p)T}$
    $\tilde{\mathbf{A}}^{i(p+1)} \leftarrow \frac{\tilde{\mathbf{A}}^{i(p+1)}}{\max\left(1, \left\|\tilde{\mathbf{A}}^{i(p+1)}\right\|_2\right)}, \forall i = 1..n$
    Compute L_S
    $\tilde{\mathbf{S}}^{(p+1)} \leftarrow \mathcal{S}_{\frac{\Lambda}{L_S}}\left(\left(\tilde{\mathbf{S}}^{(p)} + \frac{1}{L_S}\tilde{\mathbf{A}}^{(p+1)T}\left(\mathbf{X} - \tilde{\mathbf{A}}^{(p+1)}\tilde{\mathbf{S}}^{(p)} - \tilde{\mathbf{O}}^{(p)}\right)\right)\mathbf{\Phi}_\mathbf{S}^T\right)\mathbf{\Phi_S}$
    $\tilde{\mathbf{O}}^{(p+1)} \leftarrow \boldsymbol{\alpha}_{\tilde{\mathbf{O}}^{(p+1)}}\mathbf{\Phi_O}$ where $\forall j = 1..t$ and $\forall i = 1..m$:
    $\boldsymbol{\alpha}_{\tilde{\mathbf{O}}^{(p+1)}}{}_i^j \leftarrow \left(\left(\left(\mathbf{X} - \tilde{\mathbf{A}}^{(p+1)}\tilde{\mathbf{S}}^{(p+1)}\right)\mathbf{\Phi}_\mathbf{O}^T\right)_i^j \times \max\left(0, 1 - \frac{\Upsilon_i^j}{\left\|\left(\left(\mathbf{X} - \tilde{\mathbf{A}}^{(p+1)}\tilde{\mathbf{S}}^{(p+1)}\right)\mathbf{\Phi}_\mathbf{O}^T\right)^j\right\|_2}\right)\right)$
  return $\tilde{\mathbf{S}}^{(P-1)}$, $\tilde{\mathbf{A}}^{(P-1)}$, $\tilde{\mathbf{O}}^{(P-1)}$.

Limitations of the standard block coordinate minimizers. The proposed prox-linear implementation is sensitive to the setting of the parameters and the initialization, which makes the joint estimation of the regularization parameters and the components highly challenging. In practice, algorithms like GMCA for standard sparse BSS [5], are based on BCD but with additional heuristics, which play a key role to provide robustness to the initialization and an automatic setting of the parameters. If these heuristics, which are detailed in Section 3.3, yield more robust minimization procedures, they lack provable convergence. Therefore the global optimization strategy used in the tr-rGMCA algorithm is composed of two successive steps:

- The warm-up step: a solution of Problem (7) is approximated using a BCD-based algorithm with heuristics. This first step, which is described in Section 3.3, aims at

providing a robust first guess of the components as well as the parameters values for the next, provably convergent, step.

- The refinement step: The goal of this stage, which we described in Alg.1, is to provide a local stationary point of (7).

We point-out that the efficient warm-up procedure is key to the matrix-factorization problem, and prevents the computationally intensive and inefficient multi-starts method. We would like to highlight that the need for appropriate heuristics to build reliable matrix factorization procedures has also been pointed out in the framework of NMF in [24].

3.3. Warm-up procedure. In this section, we describe the so-called "warm-up"stage of the tr-rGMCA algorithm. This procedure aims at providing an approximated solution of Problem (7) as well as robustly determining the regularization parameters. The proposed strategy builds upon an appropriate choice of the variables to be updated based on either morphological or spectral diversity, that leads to the following BCD-like procedure:

- **Joint estimation of O and S based on morphological diversity.** Jointly estimating $\mathbf{O}$ and $\mathbf{S}$ for fixed $\mathbf{A}$ amounts to solving the following convex optimization problem:

$$\underset{\mathbf{S},\mathbf{O}}{\text{minimize}}\, \frac{1}{2}\left\|\mathbf{X}-\mathbf{A}\mathbf{S}-\mathbf{O}\right\|_2^2+\left\|\Lambda\odot\mathbf{S}\mathbf{\Phi}_{\mathbf{S}}^T\right\|_1+\left\|\Upsilon\odot\mathbf{O}\mathbf{\Phi}_{\mathbf{O}}^T\right\|_{2,1}, \tag{14}$$

  which we previously interpreted as a multichannel extension of Morphological Component Analysis. This step, which is detailed in Section 3.3.1, essentially exploits the morphological diversity between the outliers and the sources.

- **Joint estimation of A and S based on spectral diversity.** Updating $\mathbf{A}$ and $\mathbf{S}$ boils down to tackling the sparse BSS problem from the residual term $\mathbf{X}-\mathbf{O}$:

$$\underset{\mathbf{S},\mathbf{A}}{\text{minimize}}\, \frac{1}{2}\left\|\mathbf{X}-\mathbf{A}\mathbf{S}-\mathbf{O}\right\|_2^2+\left\|\Lambda\odot\mathbf{S}\mathbf{\Phi}_{\mathbf{S}}^T\right\|_1+\chi_{\mathbf{Y}:\left\|\mathbf{Y}^k\right\|_2\leq 1,\forall k}\left(\mathbf{A}\right). \tag{15}$$

  While being non-convex, algorithms like GMCA [5] or AMCA [4] provide efficient approximate minimization that have been shown to be robust to spurious local stationary points. This stage is described in Section 3.3.2.

The warm-up procedure alternates between these two problems to minimize (7), such as presented in Alg.2. As it will be described in the remaining of this subsection, the warm-up procedure involves key heuristics that rely on particular parameter strategies and approximations which are made to fasten the process and improve its robustness with respect to the initialization and the spurious local stationary points.

In the numerical experiment section 5, we provide a comparison between the performances of the warm-up step alone, the refinement (PALM-based) step alone, and the combination of both (tr-rGMCA), showing the robustness of the warm-up procedure as well as the benefit in term of accuracy for using the refinement step.

Algorithm 2 WarmUp Procedure

procedure WarmUp($\mathbf{X}, \tilde{\mathbf{A}}, \mathbf{\Phi_S}, \mathbf{\Phi_S}$)
  Initialize $\tilde{\mathbf{S}}^{(k=0)} \leftarrow 0$, $\tilde{\mathbf{A}}^{(k=0)} \leftarrow \tilde{\mathbf{A}}$ and $\tilde{\mathbf{O}}^{(k=0)} \leftarrow 0$.
  while $k < K$ **do**
    Set $\boldsymbol{\alpha_S}^{(i=1,k)} \leftarrow \tilde{\mathbf{S}}^{(k-1)}\mathbf{\Phi}_\mathbf{S}^T$, $\tilde{\mathbf{A}}^{(i=1,k)} \leftarrow \tilde{\mathbf{A}}^{(k-1)}$, and $\boldsymbol{\alpha}_{\mathbf{X-O}} \leftarrow \left(\mathbf{X} - \tilde{\mathbf{O}}^{(k-1)}\right)\mathbf{\Phi}_\mathbf{S}^T$
    while $i < I$ **do** ▷ Joint estimation of $\mathbf{A}$ and $\mathbf{S}$
      $\boldsymbol{\alpha_S}^{(i,k)} \leftarrow \mathcal{S}_\Lambda\left(\tilde{\mathbf{A}}^{(i-1,k)\dagger}\boldsymbol{\alpha}_{(\mathbf{X-O})}\right)$
      $\tilde{\mathbf{A}}^{(i,k)} \leftarrow \boldsymbol{\alpha}_{(\mathbf{X-O})}\boldsymbol{\alpha_S}^{(i=1,k)\dagger}$
      Decrease Λ
    Set $\tilde{\mathbf{S}}^{(k)} \leftarrow \boldsymbol{\alpha_S}^{(i=I-1,k)}\mathbf{\Phi_S}$ and $\tilde{\mathbf{A}}^{(k)} \leftarrow \tilde{\mathbf{A}}^{(I-1,k)}$
    Set $\tilde{\mathbf{S}}^{(\ell=0,j=0,k)} \leftarrow \tilde{\mathbf{S}}^{(k)}$ and $\tilde{\mathbf{O}}^{(\ell=0,j=0,k)} \leftarrow \tilde{\mathbf{O}}^{(k-1)}$
    for $\ell < L$ **do** ▷ Reweighting Procedure
      while $j < J$ **do** ▷ Joint estimation of $\mathbf{S}$ and $\mathbf{O}$
        Update $\tilde{\mathbf{S}}^{(\ell,j,k)}$ with FISTA using the proximal gradient step (12)
        Update $\tilde{\mathbf{O}}^{(\ell,j,k)}$ with the closed form (13)
      Update Λ and Υ for the reweighting procedure according to (16)
      Set $\tilde{\mathbf{S}}^{(\ell+1,0,k)} \leftarrow \tilde{\mathbf{S}}^{(\ell,J,k)}$, $\tilde{\mathbf{O}}^{(\ell+1,0,k)} \leftarrow \tilde{\mathbf{O}}^{(\ell,J,k)}$
  return $\tilde{\mathbf{A}}^{(K)}$, $\tilde{\mathbf{S}}^{(L,J,K)}$, $\tilde{\mathbf{O}}^{(L,J,K)}$.

3.3.1. Estimating O and S using the morphological diversity. For fixed $\mathbf{A}$, the outliers $\mathbf{O}$ and the sources $\mathbf{S}$ are the solutions of Problem (14). Since, for fixed sources, updating the outliers allows a closed-form expression, we opted for the BCD strategy that alternates between estimations of $\mathbf{O}$ and $\mathbf{S}$:

- **Updating the sources.** The estimation of $\mathbf{S}$ is given by $\mathbf{P}_S$ (9). As stated in Section 3.2, $\mathbf{P}_S$ can be solved with the FB algorithm: $\mathbf{S}$ is updated with the proximal gradient step eq.12 until convergence. This algorithm is also known as Iterative Soft-Thresholding (ISTA). We point out that in practice, the accelerated FISTA [2] is preferred.
- **Updating the outliers.** The estimation of $\mathbf{O}$ is given by $\mathbf{P}_O$ (10). The corresponding update with the FB algorithm is the closed form eq.13.

Parameter updates. In this subproblem, an adapted setting of the parameters Λ and Υ is important to control the leakages between the two components and so achieve a good separation between $\mathbf{AS}$ and $\mathbf{O}$.

- Reweighted scheme: The ℓ_1 and $\ell_{2,1}$ norms introduce some biases [38], which can be detrimental to the BSS problem in the presence of outliers, or at least lead to inaccurate solutions with artifacts. For this reason, a reweighted scheme is implemented [12, 38]: the values of the parameters Λ and Υ depend on the values of the estimated variables. More precisely, we will set $\Lambda = \lambda_D \mathbf{W_S}$ and $\Upsilon = \upsilon \times \mathbf{W_O}$, where $\lambda_D \in \mathbb{R}^{n\times n}$ is a diagonal matrix, whose diagonal coefficients $\{\lambda_i\}_{i=1..n}$ set the sparsity level of each

source, and $\mathbf{W_S} \in \mathbb{R}^{n \times t_{\mathbf{\Phi_S}}}$ corresponds to the varying weights. Similarly υ is a scalar setting the global sparsity level of the columns of $\mathbf{O}\mathbf{\Phi}_\mathbf{O}^T$, while $\mathbf{W_O} \in \mathbb{R}^{m \times t_{\mathbf{\Phi_O}}}$ contains the weighting parameters. Since we assume that the Gaussian noise $\mathbf{N}$ is constant from one channel to another, the parameters $\mathbf{W_O}$ do not vary from one row to another.

- Fixed parameters λ and υ Similarly to the algorithms using sparse modeling in the presence of Gaussian noise in [43], the values $\{\lambda_i\}_{i=1..n}$ are fixed to $k\sigma_i$, where σ_i are obtained with the mad of $\left(\mathbf{A}^T\left(\mathbf{X}-\mathbf{O}-\mathbf{AS}\right)\mathbf{\Phi}_\mathbf{S}^T\right)_i$.
  The value of υ is set so as to limit the impact of the remaining Gaussian noise on the estimation of $\mathbf{O}$. Outside the support of $\mathbf{O}\mathbf{\Phi}_\mathbf{O}^T$, the ℓ_2 norm of the columns of the centered Gaussian residual follows a χ-law with m degrees of freedom, whose expectation is given by $\sigma \times \sqrt{2} \times \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}$, where σ can be estimated with the value of the mad of $\left(\mathbf{X}-\mathbf{AS}-\mathbf{O}\right)\mathbf{\Phi}_\mathbf{O}^T$. The parameter υ is set to $\upsilon = k \times \sigma \times \sqrt{2} \times \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}$.
- Weights $\mathbf{W_S}$ and $\mathbf{W_O}$. At every iteration $\ell > 1$ such as in Alg.2, the parameters $\mathbf{W_S}$ and $\mathbf{W_O}$ are updated according to the current values of $\mathbf{S}$ and $\mathbf{O}$ respectively such as:

$$\mathbf{W_S} = \frac{\lambda}{\lambda + |\left(\tilde{\mathbf{S}}\mathbf{\Phi}_\mathbf{S}^T\right)|} \text{ and } \mathbf{W}_\mathbf{O}^q = \frac{\upsilon}{\upsilon + \left\|\left(\tilde{\mathbf{O}}\mathbf{\Phi}_\mathbf{O}^T\right)^q\right\|_2} \quad \forall q = 1..t. \tag{16}$$

  We point out that $\mathbf{W_S}$ and $\mathbf{W_O}$ are reset to 1 for $\ell = 1$ so as to limit the propagation of the errors and make full benefit of the new estimation of $\mathbf{A}$ by not enforcing the solutions to be similar to the previous ones.

3.3.2. Sparse BSS for the joint estimation of A and S. The joint estimation of $\mathbf{A}$ and $\mathbf{S}$ with fixed $\mathbf{O}$ amounts to perform a standard sparse BSS on the current denoised observations $\mathbf{X}-\mathbf{O}$. For that purpose, we will make use of either the GMCA [5] or the AMCA (Adaptive Morphological Component Analysis [4]) algorithm to update these variables. The algorithm AMCA, compared to GMCA, further implements an iterative weighting scheme when estimating $\mathbf{A}$. This weighting strategy aims at penalizing the samples of $\mathbf{X}-\tilde{\mathbf{O}}$ which behave as corrupted samples, and which can be traced using the sparsity level of the estimated expansion coefficients of the sources [4]. The AMCA algorithm has been used to improve the separation of $\mathbf{A}$ and $\mathbf{S}$ in the presence of outliers when no morphological diversity can help distinguishing between the sources and the outliers [14].
During the very first iterations of the warm-up stage, a large large part of the outliers is very likely to be misestimated and still present in the residual $\mathbf{X}-\tilde{\mathbf{O}}$, which will eventually hamper the unmixing process. Choosing the BSS algorithm that is the most robust to this residual will help enhancing the estimation $\mathbf{A}$. For that purpose either GMCA or AMCA will be used based on the relative choices of $\mathbf{\Phi_S}$ and $\mathbf{\Phi_O}$:

- Highly incoherent dictionaries: If $\mathbf{\Phi_O}$ and $\mathbf{\Phi_S}$ are highly incoherent, the outlier residual is likely to be dense in $\mathbf{\Phi_S}$, similarly to the case displayed in fig.3. Using the standard fast GMCA, which is robust to the presence of Gaussian noise, and more generally to dense noise, is the best choice.

- Mildly incoherent dictionaries: In this case, the algorithm AMCA should be preferred [7]. Indeed, the representations of the outliers and their residues in $\mathbf{\Phi_S}$ are likely to be mildly sparse. In that case, we showed in [14] that the AMCA algorithm provides a more robust estimate of $\mathbf{A}$.

- Additional priors on the sources: Besides the morphology of the residual of the outliers in $\mathbf{\Phi_S}$, another additive knowledge on the data may justify the use of a specific sparse BSS algorithm. For example, if the sources are correlated, the algorithm AMCA, which was originally developed to handle partially correlated sources, should be preferred to GMCA, even if the residual of the outliers is dense in $\mathbf{\Phi_S}$.

Since AMCA and GMCA only differ by this weighting scheme, we will present the warm-up procedure using GMCA. The AMCA algorithm is implemented by adding the weighting proposed in [4].

Component updates. The fast version of GMCA performs the separation directly in the transformed domain $\mathbf{\Phi_S}$. The returned results are exact if $\mathbf{\Phi_S}$ is orthonormal, and provide a good approximation if $\mathbf{\Phi_S}$ is diagonally dominant [5]. The GMCA algorithm estimates alternatively $\mathbf{A}$ and $\boldsymbol{\alpha}_\mathbf{S}$ by minimizing:

$$\underset{\mathbf{A},\boldsymbol{\alpha}_\mathbf{S}}{\text{minimize}}\, \frac{1}{2}\left\|\left((\mathbf{X}-\mathbf{O})\mathbf{\Phi}_\mathbf{S}^T - \mathbf{A}\boldsymbol{\alpha}_\mathbf{S}\right)\right\|_2^2 + \|\Lambda \odot \boldsymbol{\alpha}_\mathbf{S}\|_1 .$$

The algorithm estimates alternatively $\mathbf{A}$ and the coefficients $\boldsymbol{\alpha}_\mathbf{S}$ with projected least-squares to fasten the unmixing process [5], [4].
The corresponding updates are given in Alg.2 and further details can be found in [5], [4].

Parameter updates. The strategies used for the setting of the parameters involved in GMCA are crucial for the robustness against the noise and local minima. They are presented below:

- The values of $\Lambda = \lambda \odot \mathbf{W_S}$ plays a key role in AMCA and GMCA. In order to adopt the efficient scheme used in [5] and to limit the propagation of the errors due to a previous misestimation of $\mathbf{S}$, the weights $\mathbf{W_S}$ are set to 1 during the unmixing process.
  In [5], the authors propose a decreasing strategy for Λ. At the beginning, only the largest coefficients, which are the most discriminant for the separation, are selected. Then, the solutions are refined by decreasing the value of λ. This "*coarse to fine strategy* "[5] improves the robustness of the algorithm against local minima. In practice, an increasing number of entries is selected at every iteration. The final threshold λ_i for each $\boldsymbol{\alpha}_{\mathbf{S}i}$ is $k\sigma_i$ where σ_i corresponds to the standard deviation of the noise corrupting the coefficients of the ith source, and $k \in (1,3)$[5]. The value of σ_i, if not known, can be estimated with the value of the mad of the coefficient $\boldsymbol{\alpha}_{\mathbf{S}i}$ before the thresholding operation.

Convergence and stability of the tr-rGMCA. Ajouter un paragraphe ?

4. Numerical experiments: algorithms for comparison and performance criteria. We compare tr-rGMCA with standard robust BSS methods. These methods as well as the different

criteria used to compare the algorithms are presented in this section. The different strategies are compared first with simulated data allowing Monte-Carlo simulations (40 runs for each varying parameter). Last, they are compared on realistic simulated data from the ESA-Planck mission in the presence of additional point-sources emissions which act as outliers.

4.1. Algorithms for the comparison. Only few methods presented in the literature can handle the considered problem. Most of these methods require additional assumptions, which will not be always valid in the following experiments. In this section, we present the selected strategies for the comparison explaining in which experiments they will be used.

Proposed optimization strategy. In order to highlight the robustness of the proposed minimization strategy, we will compare it with the following other implementations:

- **Oracle with A known.** In this case, we assume that $\mathbf{A}$ is known, and we separate $\mathbf{O}$ from $\mathbf{AS}$ using the morphological diversity between the two components:

$$\underset{\mathbf{O},\mathbf{S}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{X} - \mathbf{AS} - \mathbf{O}\|_2^2 + \left\|\Lambda \odot \mathbf{S}\mathbf{\Phi}_{\mathbf{S}}^T\right\|_1 + \left\|\Upsilon \odot \mathbf{O}\mathbf{\Phi}_{\mathbf{O}}^T\right\|_{2,1}.$$

  The difference between these results and the ones of tr-rGMCA illustrates the loss of accuracy led by the blind unmixing process.
- **The PALM procedure only.** In order to underline the advantage of using the initialization procedure, we also minimize 7 using only the refinement step in Alg.1. Since a reweighted procedure is implemented in tr-rGMCA, the refinement procedure is run three times: first, it is initialized with null $\mathbf{S}$ and $\mathbf{O}$ and the matrix $\mathbf{A}$ used for tr-rGMCA[1], and for the second and third times, the regularization parameters are updated given the current estimates of $\mathbf{S}$ and $\mathbf{O}$ with the weighting procedure eq. (16)), see Alg.**??**.

Algorithm 3 Reweighted PALM

procedure Reweighted PALM($\mathbf{X}, \tilde{\mathbf{A}}, \mathbf{\Phi}_{\mathbf{S}}, \mathbf{\Phi}_{\mathbf{S}}$)
  Initialize $\tilde{\mathbf{S}}^{(k=0)} \leftarrow 0$, $\tilde{\mathbf{A}}^{(k=0)} \leftarrow \tilde{\mathbf{A}}$ and $\tilde{\mathbf{O}}^{(k=0)} \leftarrow 0$.
  for $k < 3$ **do** ▷ Reweighting Procedure
    $\tilde{\mathbf{S}}^{(k)}, \tilde{\mathbf{A}}^{(k)}, \tilde{\mathbf{O}}^{(k)} \leftarrow$ PALM($\mathbf{X}, \tilde{\mathbf{S}}^{(k-1)}, \mathbf{\Phi}_{\mathbf{S}}, \tilde{\mathbf{O}}^{(k-1)}, \mathbf{\Phi}_{\mathbf{O}}, \tilde{\mathbf{A}}^{(k-1)}, \Lambda, \Upsilon$)
    Update Λ and Υ for the reweighting procedure according to (16)
  return $\tilde{\mathbf{A}}^{(2)}$, $\tilde{\mathbf{S}}^{2)}$, $\tilde{\mathbf{O}}^{(2)}$.

- **The warm-up step only.** The intermediate performances, obtained by the initialization step only, will be also displayed. A difference between these results and the PALM procedure would bring out the robustness of this initialization step, and the dissimilarity with the all process tr-rGMCA would show the gain of using a more precise refinement step.

[1]The mixing matrix is initialized with PCA on $\mathbf{X}$, for all algorithms

Methods used for the comparisons.

- **The combination Outlier Pursuit (OP)+GMCA.** The outliers are first estimated by applying the Outlier Pursuit algorithm [49] on $\mathbf{X}\mathbf{\Phi}_{\mathbf{O}}^T$, eq.17. Then the algorithm GMCA [5] is applied on the denoised observations $\left(\mathbf{X}-\tilde{\mathbf{O}}\right)$, eq.18:

$$i)\tilde{\mathbf{O}},\tilde{\mathbf{L}} \leftarrow \underset{\mathbf{O},\mathbf{L}:\mathbf{X}=\mathbf{O}+\mathbf{L}}{\text{argmin}} \left\|\mathbf{L}\mathbf{\Phi}_{\mathbf{O}}^T\right\|_* + \lambda \left\|\mathbf{O}\mathbf{\Phi}_{\mathbf{O}}^T\right\|_{2,1} \tag{17}$$

$$ii)\tilde{\mathbf{A}},\tilde{\mathbf{S}} \leftarrow \underset{\mathbf{A},\mathbf{S}}{\text{minimize}} \frac{1}{2}\left\|\tilde{\mathbf{L}}-\mathbf{A}\mathbf{S}\right\|_2^2 + \left\|\Lambda \odot \mathbf{S}\mathbf{\Phi}_{\mathbf{S}}^T\right\|_1 \tag{18}$$

  This strategy requires the term $\mathbf{AS}$ to be low-rank, and thus, it will only be used when $m > n$. Given that the value of λ proposed in [49] does not return satisfactory results, we choose to tune its value: we select the best $\tilde{\mathbf{A}}$ among the ones obtained from GMCA after the Outlier Pursuit for which we set the parameter λ between $\frac{1}{5\sqrt{t}}$ and $\frac{10}{\sqrt{t}}$ with a step-size of $\frac{1}{5\sqrt{t}}$.
- **The rNMF algorithm [21].** This method was initially proposed for robust unmixing of terrestrial hyperspectral images. It assumes that the components are non-negative, and that the sources samples lie in the simplex with almost pure pixels. This last assumption is not valid in the following experiments, and we will instead assume that the columns of $\mathbf{A}$ are normalized:

$$\tilde{\mathbf{A}},\tilde{\mathbf{S}},\tilde{\mathbf{O}} \leftarrow \underset{\mathbf{A}\geq 0,\mathbf{S}\geq 0,\mathbf{O}\geq 0}{\text{minimize}} \frac{1}{2}\left\|\mathbf{X}-\mathbf{A}\mathbf{S}-\mathbf{O}\right\|_2^2 + \beta\left\|\mathbf{O}\right\|_{2,1} + \chi_{\mathbf{Y}:\|\mathbf{Y}\|_2\leq 1}(\mathbf{A}).$$

  This method will be used in the experiments of Section 6, in which the components are all non-negative and the outliers sparse in the direct domain $\mathbf{\Phi}_{\mathbf{O}} = \mathbf{I}$. All the conditions required for the rNMF to be efficient will not be valid.
- **ICA based on a β divergence minimization [32]** [2]. This ICA-based method looks for an unmixing matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ such that the corresponding sources $\tilde{\mathbf{S}} = \mathbf{BX}$ are mutually independent. The independence of the sources $\tilde{\mathbf{S}}$ is measured with the β-divergence $\mathbb{D}_\beta$ between the product of their marginal $\prod_{i=1}^n p_{\mathbf{S}}(\tilde{\mathbf{S}}_i)$ and their joint distribution $p_{\mathbf{S}}(\tilde{\mathbf{S}})$, which is null if and only if the sources are independent. The cost function to be minimized is given by:

$$\underset{\mathbf{B}:\tilde{\mathbf{S}}=\mathbf{BX}}{\text{minimize}}\, \mathbb{D}_\beta(p_{\mathbf{S}}(\tilde{\mathbf{S}})\|\prod_{i=1}^n p_{\mathbf{S}}(\tilde{\mathbf{S}}_i))$$

  We will only use this method when $m = n$ since otherwise, a dimension reduction technique is needed (and is challenging in the presence of outliers). Besides, it only returns $\mathbf{A}$, and thus, does not perform the separation between the outliers and the sources contribution.
  In contrast to the other methods, a strong morphological diversity makes the unmixing more challenging for this method. Indeed, it should be performed in a domain in which few samples are corrupted, and so in $\mathbf{\Phi}_{\mathbf{O}}$. However, if the morphological diversity is

[2] python implementation from [23]

strong, then the expansion coefficients of the sources in $\mathbf{\Phi_O}$ are highly non-sparse (see for example fig.3(c), the sources coefficients in $\mathbf{\Phi_O}$ almost follow a Gaussian distribution): this is difficult to handle for ICA-based methods. On the other hand, if $\mathbf{\Phi_S}$ and $\mathbf{\Phi_O}$ are not highly incoherent, then the outliers are likely to not corrupt all the samples in $\mathbf{\Phi_S}$. It is then preferable to perform the minimization in $\mathbf{\Phi_S}$ since the sources are better represented and the outliers do not corrupt all samples.

Last, setting the value of β is challenging in practice. We select the best $\mathbf{A}$ for the 20 preselected values of β, starting from 10^{-4} to 0.85.

- **GMCA [5].** This a standard sparse BSS algorithm. It will be performed on $\mathbf{X\Phi_S^T}$. Its results illustrate the sensitivity of the standard (non-robust) methods to the outliers.
- **The combination MCA [20]+GMCA.** Similarly to the combination OP+GMCA, the outliers are first discarded from the observations with MCA, Problem 19, and the unmixing is then performed on the cleaned data with GMCA 20:

$$\text{(19)} \quad \forall i = 1...m,\ \tilde{\mathbf{O}}_i, \tilde{\mathbf{L}}_i \leftarrow \underset{\mathbf{O}_i,\mathbf{L}_i}{\text{minimize}} \frac{1}{2} \|\mathbf{X}_i - \mathbf{L}_i - \mathbf{O}_i\|_2^2 + \alpha \left\|\mathbf{L}_i \mathbf{\Phi}_\mathbf{S}^T\right\|_0 + \nu \left\|\mathbf{O}_i \mathbf{\Phi}_\mathbf{O}^T\right\|_0$$

$$\text{(20)} \quad \tilde{\mathbf{A}}, \tilde{\mathbf{S}} \leftarrow \underset{\mathbf{A},\mathbf{S}}{\text{minimize}} \frac{1}{2} \left\|(\mathbf{X} - \tilde{\mathbf{O}}) - \mathbf{AS}\right\|_2^2 + \left\|\Lambda \odot \mathbf{S}\mathbf{\Phi}_\mathbf{S}^T\right\|_1 + \chi_{\mathbf{Y}:\|\mathbf{Y}\|_2 \leq 1}(\mathbf{A})$$

Instead of using the spectral diversity such as done by the OP algorithm, this combination only exploits the morphological diversity to discard the outliers. It is indeed possible to separate $\mathbf{AS}$ from $\mathbf{O}$, without regarding the ratio $\frac{n}{m}$, as long as $\mathbf{AS}$ is sparse in $\mathbf{\Phi_S}$. We point out that this hypothesis can be valid only in the presence of a small number of sources. Besides, this approach does not take into account the clustered, structural aspect of the product $\mathbf{AS}$.

4.2. Performance criteria. In this section, we present the different criteria used to compare the algorithms. In the context of robust BSS, they should assess the unmixing of the sources (recovery of $\mathbf{A}$), the separation between the outliers and the sources as well as the reliability of the separation (especially because the problem is not convex).

Unmixing.

- For each recovered $\tilde{\mathbf{A}}$, the global quantity $\Delta_A = -10 \log 10 \left(\frac{\|\tilde{\mathbf{A}}^\dagger \mathbf{A} - \mathbf{I}\|_1}{n^2} \right)$ is computed [5]. A large value denotes a good estimation of $\mathbf{A}$.
- For each recovered $\tilde{\mathbf{A}}$, the maximal angle between the estimated and actual columns of $\mathbf{A}$ is computed: $\max_{j=1..n} \arccos\langle \tilde{\mathbf{A}}^j, \mathbf{A}^j \rangle$ (in degree).

  For every considered parameter, we sum the number of runs for which an algorithm has returned a mixing matrix whose maximal angle is smaller than 5 degrees. This quantity, normalized to 1, provides a good indicator of the reliability of the algorithms.

Estimation of the sources and outliers.

- In [48], the authors decompose each retrieved audio source s as the sum:

$$s = s_{target} + s_{interference} + s_{noise} + s_{artifacts}.$$

A similar decomposition can be employed for more general signals and images [38], where s_{target} denotes the projection of the retrieved source on the sought-after one, $s_{interference}$ the residue due to the interferences with the other sources, s_{noise} accounts for the part due to the presence of noise (the outliers in our case), and last, $s_{artifacts}$, represents the remaining artifacts (coming from the leakages from $\mathbf{S}$ towards the estimated outliers, and the bias). This decomposition is used to derive the following indicators [48]:

- Signal to Distortion Ratio $SDR(s) = 20\log\left(\frac{\|s_{target}\|_2}{\|s_{interference}+s_{noise}+s_{artifacts}\|_2}\right)$.
- Signal to Interference Ratio $SIR(s) = 20\log\left(\frac{\|s_{target}\|_2}{\|s_{interference}\|_2}\right)$.
- Signal to Noise Ratio $SNR(s) = 20\log\left(\frac{\|s_{target}+s_{interference}\|_2}{\|s_{noise}\|_2}\right)$.
- Signal to Artifact Ratio $SAR(s) = 20\log\left(\frac{\|s_{target}+s_{interference}+s_{noise}\|_2}{\|s_{artifacts}\|_2}\right)$.

In Section 5, we will only display the median over the n sources of the SDR: it provides a global criterion on the precision of the source estimation. In Section 6, the medians as well as the minima for the n sources of the SDR, SAR, SIR and SNR will be displayed, so as to describe more precisely the obtained estimations.

- The sources can be erroneously estimated whereas the outliers and $\mathbf{AS}$ are correctly estimated (for the Frobenius norm). To measure the quality of the separation between $\mathbf{AS}$ and the outliers, the two components of interest for MCA and OP, we also compute the following metric for the outliers: $\mathbf{O}_{SE} = -10\log\frac{\|\tilde{\mathbf{O}}-\mathbf{O}\|_2}{\|\mathbf{O}\|_2}$, where $\mathbf{O}$ denotes the initial outliers, $\tilde{\mathbf{O}}$ the estimated ones.

5. 1D Simulations.

We start by comparing the different strategies on 1D data allowing Monte-Carlo simulation, with varying parameters. For this purpose, we will generate two kinds of data sets which are described in the next part.

5.1. Dataset.

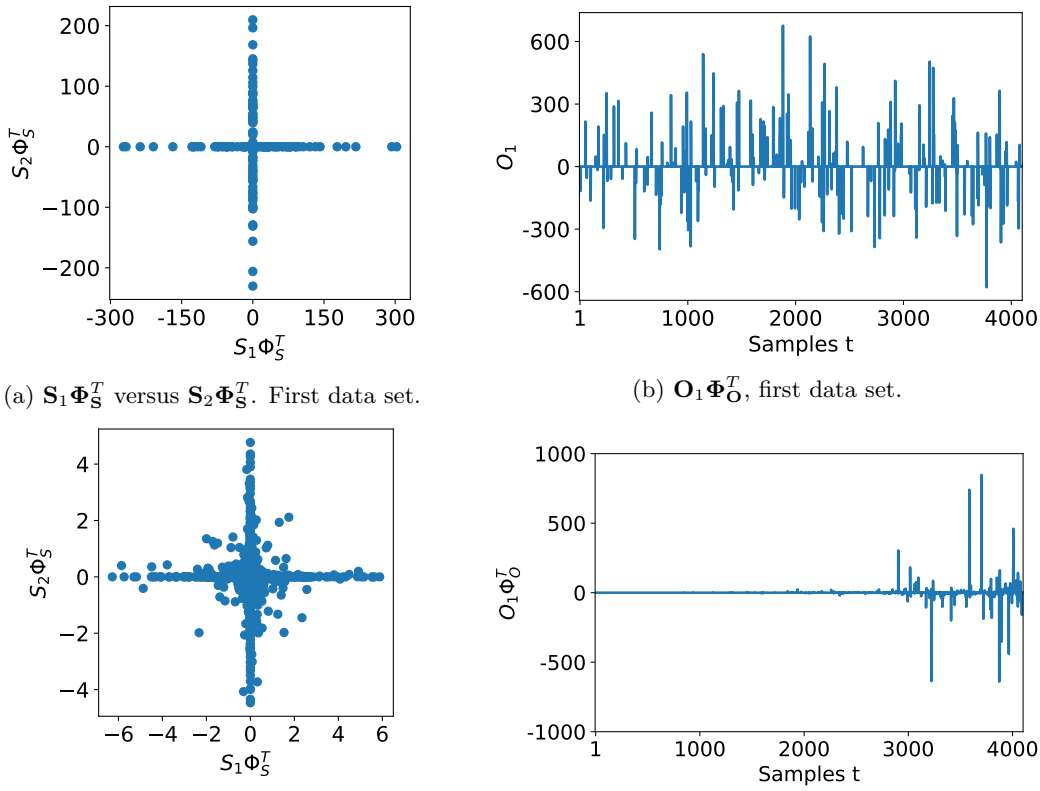
- **Dataset 1:** we consider n sources whose expansion coefficients are exactly sparse in DCT. They are drawn from a Bernoulli-Gaussian law, with an activation parameter of 5% and a standard deviation of 100. These sources are mixed into m observations, which are corrupted by outliers and an additive Gaussian noise. The outliers are sparse in the direct domain ($\mathbf{\Phi_O} = \mathbf{I}$). The support of the active columns of $\mathbf{O}$ follow a Bernoulli law, with varying activation rates fig.5. The amplitude of the active entries are drawn from a centered Gaussian distribution, with a standard deviation equal to $100 \times \frac{8}{m}$ (so that for the two considered numbers of observations $m = 8$ and $m = 40$, it will remain quite constant relatively to the amplitude of $\mathbf{AS}$). The number of samples is fixed to 4096. For the two data-sets, the entries of the mixing matrix are drawn from a Gaussian distribution and the columns of $\mathbf{A}$ are then normalized for the ℓ_2 norm. Besides, $\mathbf{A}$ is generated so as to have a condition number smaller than 100.

- **Dataset 2:** this is a more realistic setting, with a same number of samples $t = 4096$. The sources are first generated from a Bernoulli Gaussian law in the direct domain, with an activation rate of 2% and a standard deviation of 100. The sources are then convolved with a Laplacian kernel (FWHM equal to 20), fig.8. They can be sparsely represented using redundant 1D wavelets [42], fig.4c. The outliers are generated so as to correspond to a high frequency structured noise- approximately column sparse in the DCT domain. First, we generated a 1×4096 vector whose entries are drawn from a generalized Gaussian distribution, centered, with an unit variance and scale parameter 0.1. In order to obtain a high frequency texture, the amplitude of the DCT coefficients are scaled (from 10^{-4} for the lowest frequency to 1 for the highest one, with a logarithmic range), and the lowest 500 coefficients are manually set to 0. Last, this vector is multiplied (dot-wise) by a matrix generated from a Gaussian distribution, whose columns are normalized for the ℓ_2 norm, so that $\mathbf{O}\mathbf{\Phi}_{\mathbf{O}}^T$ is approximately column sparse, fig.4d.

The first dataset is almost ideal since the expansion coefficients are exactly sparse and the mutual coherence between the DCT and the direct domain is very low - see for instance fig. 3 which based on this setting. On the other hand, the second one is more realistic: the expansion coefficients are approximately sparse and the mutual coherence between the wavelets and the DCT is larger than the one between DCT and the direct domain, fig.4.

5.2. 1D Monte-Carlo simulations - Optimization strategy. In this first set of experiments using the first data-setting for 8 sources and 8 observations, we consider an easy setting to compare different optimization strategies which can be used to minimize (7). The SNR for the additive Gaussian noise is set to 30dB.

First, one can notice in Fig.5 that in the presence of very few outliers (percentage of corrupted columns equal to 1%), the different strategies perform similarly in term of precision and reliability. Moreover, their corresponding values of the SDR Fig.5 is also close to the one obtained by the oracle: the unmixing task does not hinder the estimation of the sources. However, in the presence of numerous outliers, some disparities appear: the different strategies do not perform similarly and as well as the oracle. The PALM implementation (refinement step of tr-rGMCA) is more precise than the initialization step for the unmixing (Δ_A has larger values Fig.5), but it is not as robust: except when there are only very few outliers, it cannot recover $\mathbf{A}$ for all the runs, contrary to the initialization step (with a percentage smaller than 30%) Fig.5. However, adding the refinement step after the initialization step (the proposed strategy for tr-rGMCA) allows a significant gain in term of precision: all the values of the performance indicators are higher with tr-rGMCA than with the initialization step only. Moreover, the SDR and the error for the outliers Fig.5 obtained with tr-rGMCA are very closed to the ones of the oracle: the unmixing of tr-rGMCA is robust and does not deteriorate the estimation of the sources while the percentage of corrupted columns is smaller than 30%. On the overall, tr-rGMCA is almost not influenced by the percentage of corrupted columns while this one is smaller than 30%. However it quickly fails, similarly to the oracle, in the presence of a larger percentage. Even if the dictionary chosen for $\mathbf{O}$ is not the most adapted



(a) $\mathbf{S}_1\mathbf{\Phi}_\mathbf{S}^T$ versus $\mathbf{S}_2\mathbf{\Phi}_\mathbf{S}^T$. First data set.

(b) $\mathbf{O}_1\mathbf{\Phi}_\mathbf{O}^T$, first data set.

(c) $\mathbf{S}_1\mathbf{\Phi}_\mathbf{S}^T$ versus $\mathbf{S}_2\mathbf{\Phi}_\mathbf{S}^T$. Second data set.

(d) $\mathbf{O}_1\mathbf{\Phi}_\mathbf{O}^T$, second data set.

Figure 4: Scatter plots of two expansion coefficients of the sources (left), and illustrations of the expansion coefficients of the outliers (right). The top row corresponds to the first data-set, with exactly sparse coefficients, and the second row, to the second data-set, with compressible signals.

one ($\mathbf{O}$ does not have a very sparse representation), the separation between the outliers and the source contribution can be good as long as the components have sparser representation in their associated dictionary than in the other one.
These results support the proposed strategy used for tr-rGMCA. In the following, only the results obtained by the oracle and tr-rGMCA will be displayed.

5.3. 1D Monte-Carlo simulations - Comparison in the determined case. Only few methods able to handle the presence of outliers in the determined case are present in the literature. We propose to compare these methods with tr-rGMCA in this challenging setting.

Influence of the percentage of corrupted columns - 2 Sources. In this experiment, the data are generated with the first data-set with 2 sources and 2 observations. The SNR, for the

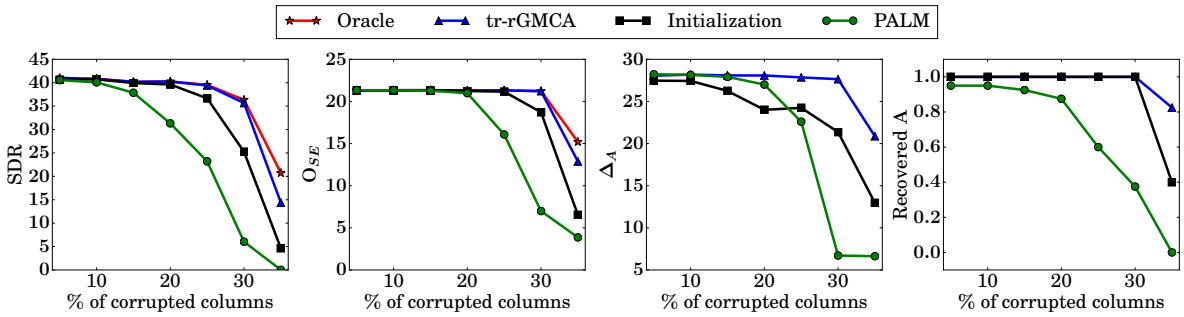


Figure 5: Performance indicators for a varying percentage of corrupted columns in the determined case for different optimization strategies.

Gaussian noise, is set to 60dB. In the determined setting, one can envisage using the minimiza-

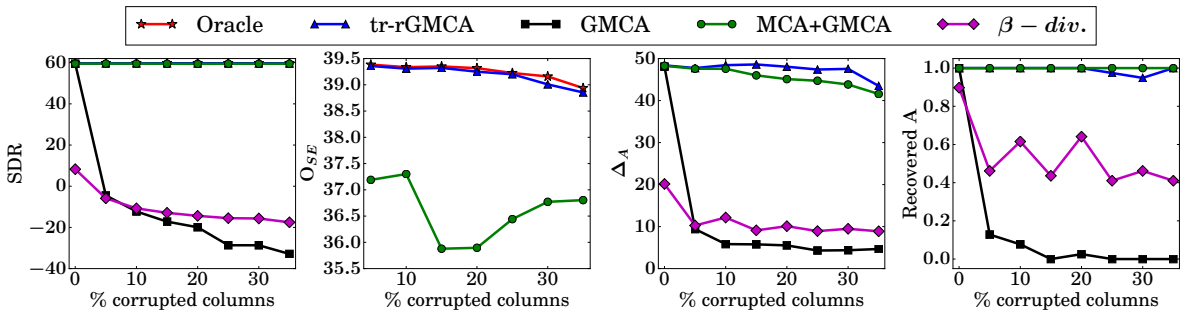


Figure 6: Performance indicators for a varying percentage of corrupted columns in the determined case for 2 sources.

tion of the β-divergence, and the combination MCA+GMCA (which can also be used in the over-determined setting). The results obtained by the minimization of the β-divergence, fig.6 are better than the ones of GMCA but not as reliable or as precise as the ones of tr-rGMCA or MCA+GMCA. However, we explained in the presentation of the different methods used for the experiments, that this setting is challenging for the minimization of the β-divergence. Besides, the parameter β needs to be finely tuned, and we only tried 20 different values for this parameter.

The second comment that can be made regarding fig.6, is on the impressive performances of the combination MCA+GMCA which performs very similarly to tr-rGMCA and the so called oracle. We will see in the next experiments that the combination MCA+GMCA is nonetheless not able to handle the presence of a larger number of sources.

Influence of the amplitude of the outliers - 8 Sources . The data are generated from the second data-setting. We consider that 8 sources have been mixed into 8 observations. The SNR for the Gaussian noise is set to 50dB. In this experiment, we observe the influence of

the amplitude of the outliers. For this purpose, we define the SOR (signal to outlier ratio), similarly to the SNR: $SOR = 20 \log \frac{\|\mathbf{AS}\|_2}{\|\mathbf{O}\|_2}$. The support of the outliers remains constant for a given run, and only their amplitude is modified, by setting the SOR according to the value of the x-axis of Fig.7.

On the overall, the values of the different performance indicators are smaller than with

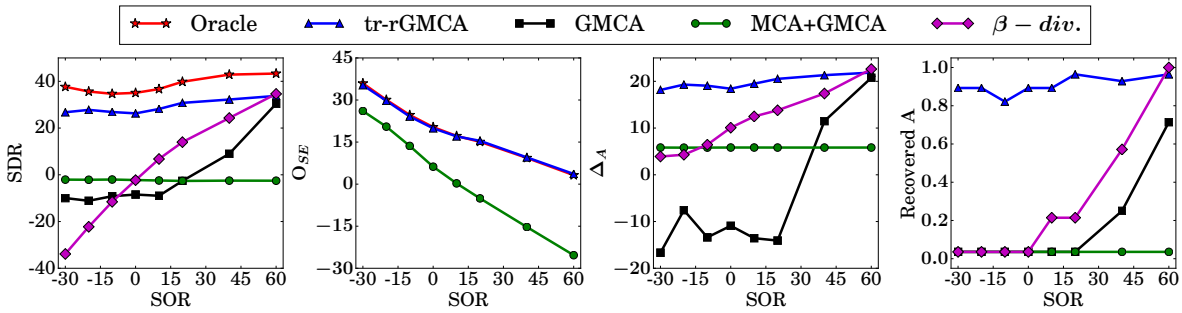
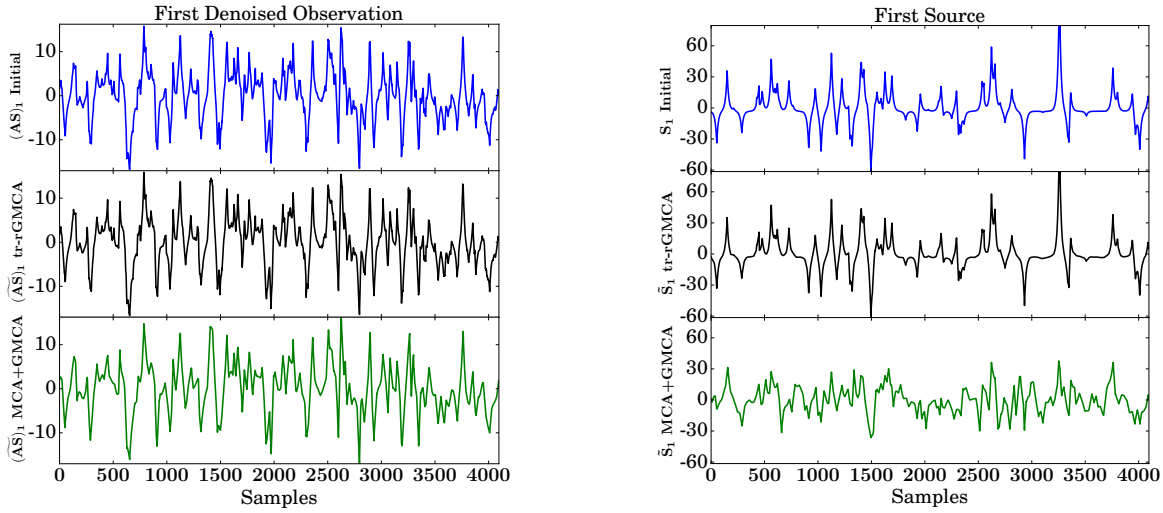


Figure 7: Performance indicators for a varying amplitude of the outliers in the determined case for 8 sources.

the first data-set: the second data-set, more realistic, is indeed more complicated. More specifically, one can note the significant gap for the SDR between the oracle and the other methods fig.7, whereas the outlier estimations have a similar precision fig.7: the additional unmixing clearly affects the results. The discrepancy between the minimization of the β-divergence and tr-rGMCA is reduced in this setting: the minimization of the β-divergence is performed in $\mathbf{\Phi_S}$, which is favorable to the unmixing.
With this data set and this number of sources, MCA fails to separate the outliers from the source contributions, and the consecutive GMCA returns erroneous solutions. It fails because the component $\mathbf{AS}$ is not sparse enough in $\mathbf{\Phi_S}$ (the number of sources is too large), and that it does not take into account the structure, the clustered aspect of the product $\mathbf{AS}$. This is illustrated in Fig.8: the estimation of $\mathbf{AS}$ obtained by MCA+GMCA is fair, but the resulting sources are clearly not correctly estimated. On the other side, the proposed tr-rGMCA is robust to outliers having a large amplitude, at least, much more than the standard BSS method GMCA.

5.4. 1D Monte-Carlo simulations - Comparison in the over-determined case. In Section 2, we underline the importance of the ratio $\frac{m}{n}$ in robust BSS. To illustrate it, we vary the number of observations, for 6 sources, with the two data-settings. The SNR is fixed to 50 dB and the SOR to -10dB for the first data set and 10dB for the second. Besides, the condition number of $\mathbf{A}$, which plays a crucial role in robust BSS, is very likely to decrease with an increasing m. In order to limit the influence of this parameter, the condition number of $\mathbf{A}$ is limited to 5.

First data-set. We start with the first data-set. The results obtained by the different methods are improved if $m \gg n$, fig.9. Given that the outliers in $\mathbf{\Phi_S}$ are broadly distributed, they



(a) Restoration of the first observation.

(b) Estimation of the first source.

Figure 8: Illustrations of the estimated signals for a SOR equal to -10dB. On the left, restoration of the first observation, on the right, estimation of the first source. In blue, the initial signals, in black, the ones recovered by tr-rGMCA and in green by the combination MCA+GMCA.

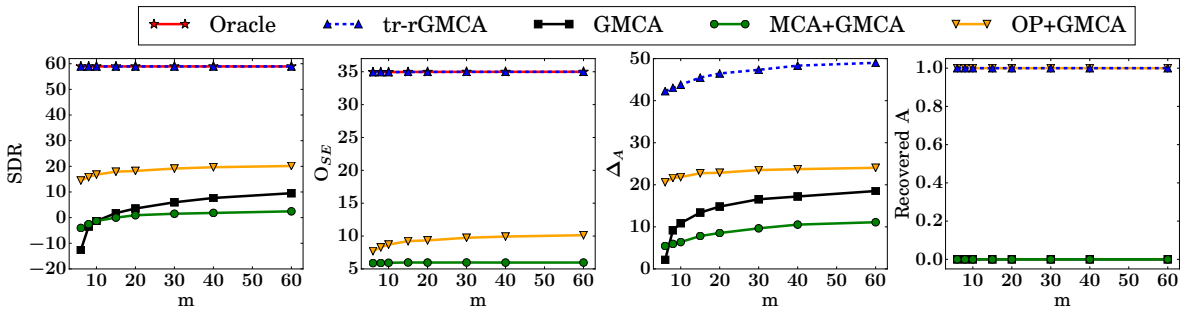


Figure 9: Performance indicators for a varying number of observations, m, for 6 sources for the first data set

behave similarly to an additive Gaussian noise with a large variance. Most of the methods used for the comparison used GMCA which is robust to the presence of a large Gaussian noise thanks to the thresholding operator whose threshold value varies according to the current noise level. However, this large threshold value leads to the presence of artifacts and biased source coefficients. When the number of observations becomes large, the projection of outliers in the span of $\mathbf{A}$ has a smaller energy, and so, the corresponding apparent noise level becomes also smaller: the artifacts become also smaller, and both $\mathbf{A}$ and $\mathbf{S}$ are more accurate. That is why, most of the methods are able to estimate $\mathbf{A}$, and $\mathbf{S}$ fairly when $m \gg n$.

It can also be noticed that even if m is close to n, the combination Outlier Pursuit (OP) + GMCA is able to retrieve $\mathbf{A}$, while GMCA alone cannot. The sources and the outliers are not precisely retrieved, but the results are the second best after tr-rGMCA. With the strong morphological diversity, the outliers are very sparse in $\mathbf{\Phi_O}$ while the source contribution is very dense: sparsity is discriminative enough, and OP can discard a part of the outliers. Removing the largest outlier contribution is sufficient, since this data set is very favorable to GMCA.

Second data-set. The different methods are on the whole less performing with the second data set, even if an improvement is also noticeable if $m \gg n$, especially for the combination OP+GMCA fig.10. The proposed tr-rGMCA is the only method able to estimate precisely

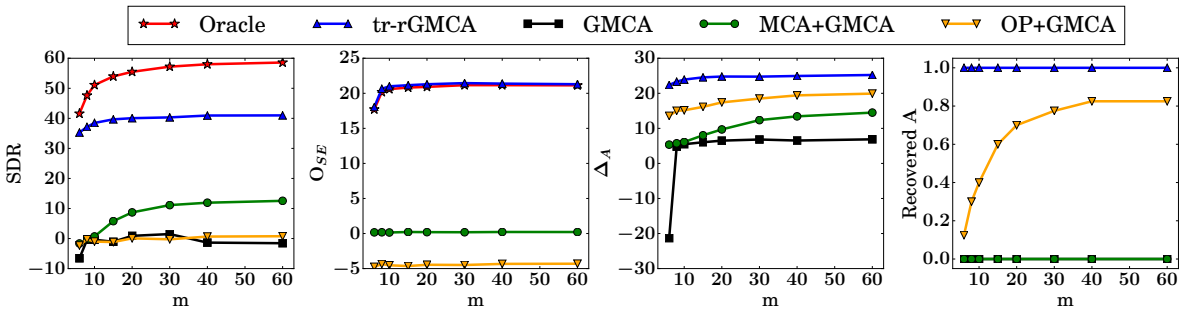


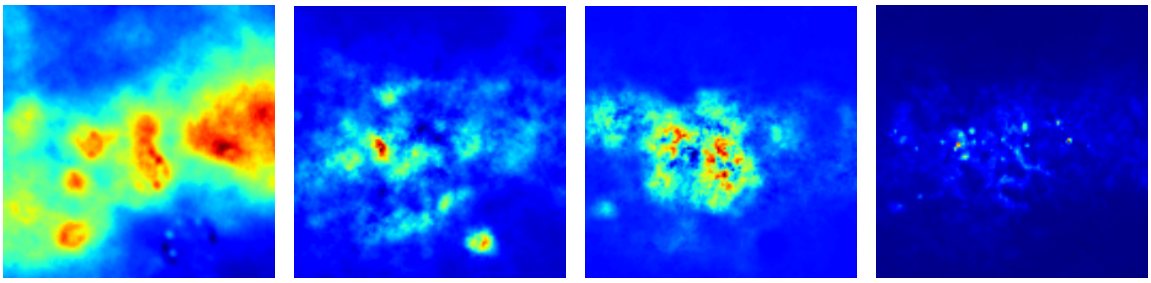
Figure 10: Performance indicators for a varying number of observations, m, for 6 sources.

and reliably the three variables, including when the number of sources is close to the number of observations. The combination MCA+GMCA struggles to solve the problem because the number of sources is too large ($\mathbf{AS}$ is not sparse enough). The algorithm Outlier Pursuit (OP) cannot identify precisely the outliers and the term $\mathbf{AS}$, but can discard efficiently the part of the outliers that are detrimental for the unmixing (the results obtained for $\mathbf{A}$ are fair).

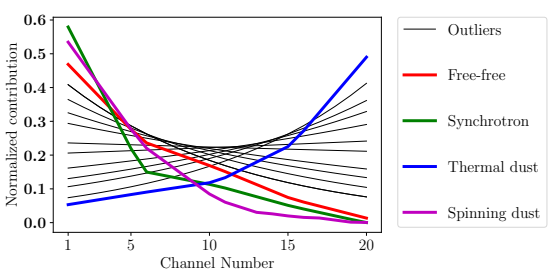
6. Application to simulated astrophysical data. In the field of astrophysics, BSS plays a central role to analyse the data from now widespread multi-wavelength instruments. More particularly, it made possible the estimation of high accuracy estimates of the Cosmological Microwave Background (CMB) from multi-wavelength microwave Planck data [7, 30]. In this context, each observation measures a linear combination of various components of our Universe. These emissions are essentially dominated by galactic components: the free-free emission, galactic synchrotron emission, spinning dust and thermal dust emissions – see [18] for more details about astrophysical microwave emissions.
However, the presence of point-source emissions and spectral variabilities of some of the galactic foreground emissions are not precisely described by the standard linear mixture model. That is why most of the component separation methods only seek for a partial CMB map, in which the galactic center and the point source emissions of known locations are masked. Since each point source has a specific spectral signature, they cannot be modeled as individual components and are rather considered as outliers. We therefore propose applying the tr-rGMCA algorithm to robustly estimate the galactic emissions (once the CMB is estimated

and its contribution discarded from the observations) in the presence of unknown point source emissions.

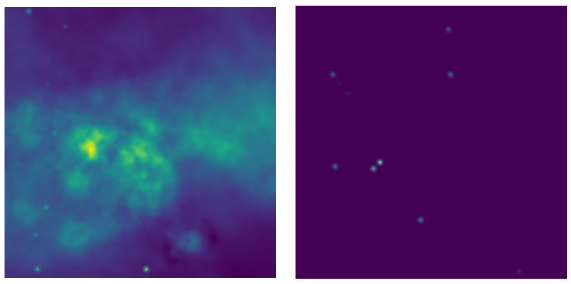
6.1. Simulated data. In the following, we simulate 20 realistic CMB-free observations $\mathbf{X} \in \mathbb{R}^{20\times 16384}$ (each image of size 128×128 is vectorized) in the microwave range at the proximity of the galactic center, which have been produced using the Planck Sky Model [18]. These observations correspond to the mixture of 4 galactic emissions, namely, synchrotron, spin dust, free-free, and thermal dust, so that $\mathbf{S} \in \mathbb{R}^{4\times 16384}$. Since the rank of $\mathbf{AS}$ is 4 and the number of observations is fixed to 20, it will make sense to apply as well separation methods that assume the low-rankness of the sources' contribution. The signal-to-noise ratio (for the Gaussian noise $\mathbf{N}$) is set to 60 dB. Ten extra point source emissions with different emission laws are added, $\mathbf{O} \in \mathbb{R}^{20\times 16384}$: $\mathbf{O}$ is composed of 10 different active columns $\{t_k\}_{k=1..10}$. These point sources, modeled as Diracs, are then convolved with a same Gaussian kernel $\mathbf{h} \in \mathbf{R}^{1\times 16384}$ with varying width w, accounting for the point spread function (beam) of the instrument fig.11: $\mathbf{X}_i = \mathbf{A}_i\mathbf{S} + \mathbf{h} * \mathbf{O}_i + \mathbf{N}_i$. In the following, we will note $\mathbf{H} \circledast \mathbf{O}$, this observation-wise convolution.



(a) Synchrotron emission. (b) Spin dust emission. (c) Free-free emission. (d) Thermal dust emission.



(e) Normalized spectra of the emissions.



(f) $\mathbf{X}_1$. (g) $\mathbf{O}_1$.

Figure 11: Top row: the 4 initial emissions. Second row: (left) normalized spectra of the emissions (*i.e.* columns of $\mathbf{A}$ and active columns of $\mathbf{O}$), and then illustrations of the first observation and corresponding outliers, for a width of the kernel equal to 1.

6.2. Upgrades of tr-rGMCA. In contrast to the tr-rGMCA algorithm we used so far, additional properties can be accounted for in the separation:

- Non-negativity of the mixing matrix and the sources. In this application, all the variables are non-negative. Taking into account non-negativity of $\mathbf{S}$ and $\mathbf{O}$ is particularly efficient to limit the leakages and artifacts between the two contributions. Non-negativity is constrained in the version of the tr-rGMCA algorithm that we used in the next experiments.
- Convolutive model for the point sources. The outliers are sparse in the direct domain. However, each one is perfectly described as the convolution of the instrument PSF and a Dirac with unknown position and amplitude. Therefore, the tr-rGMCA algorithm is extended so as to account for this convolutive model. We underline that even if the outliers are sparse in the direct domain, the morphological diversity occurs between the Gaussian kernel and $\mathbf{\Phi_S}$. In the following, by an abuse of notation, we designate by $\mathbf{\Phi_O}$ the set of all possible shifted Gaussian kernels: the observed point source emissions are sparsely represented in $\mathbf{\Phi_O}$ - once deconvolved with the Gaussian kernel.

Consequently, we slightly modify the cost function of tr-rGMCA as follows:

$$\underset{\mathbf{A},\mathbf{S}\geq 0,\mathbf{O}\geq 0}{\text{minimize}} \frac{1}{2}\left\|\mathbf{X}-\mathbf{AS}-\mathbf{H}\circledast\mathbf{O}\right\|_2^2+\left\|\Upsilon\odot\mathbf{O}\right\|_{2,1}+\left\|\Lambda\odot\mathbf{S}\mathbf{\Phi}_{\mathbf{S}}^T\right\|_1+\chi_{\mathbf{Y}:\left\|\mathbf{Y}^k\right\|_2\leq 1,\forall k}\left(\mathbf{A}\right). \tag{21}$$

The non-negativity constraints and the deconvolution are taken into account during the joint estimation of $\mathbf{O}$ and $\mathbf{S}$ of the warm-up procedure as well as the refinement step. This does not change the structure of the algorithm and the BSS method that is used to estimate jointly $\mathbf{A}$ and $\mathbf{S}$ (AMCA). Only the updates of $\mathbf{O}$ and $\mathbf{S}$ are changed during their joint estimation in the warm-up and the PALM algorithm.

The cost function of the subproblem associated with the update of $\mathbf{O}$ is composed of one differentiable term, with a Lipschitz gradient, and a regularization term (non-negativity and $\ell_{2,1}$ norm) whose proximal operator has a closed form: the update of $\mathbf{O}$ can be efficiently tackled using the FB algorithm.

On the other hand, the minimization problem associated with the update of $\mathbf{S}$ is also composed of a differentiable term with Lipschitz gradient, and two regularization terms (non-negativity in the direct domain and sparsity in a transformed domain such as in [38]), having both explicit proximal operators. This subproblem is well handled by the Generalized Forward Backward Splitting algorithm [37].

6.3. Experiments.

6.3.1. A challenging setting. First, we underline that the proposed problem is particularly difficult to tackle:

- It has first been noticed that the large scales of these astrophysical sources are partially correlated [4], which dramatically hampers the performances of standard BSS. This is precisely for this type of sources that the AMCA algorithm [4] has been designed. Therefore, the AMCA algorithm will be used in the warm-up stage to provide robustness with respect to these partially correlations.

- Some features of the thermal dust emission 11d have morphologies that are close to the one of the outliers 1b. The dictionary $\mathbf{\Phi_S}$ should be chosen so that all the sources are well represented, and also so that $\mathbf{\Phi_O}$ and $\mathbf{\Phi_S}$ are incoherent. More precisely,

the astrophysical sources admit an approximately sparse representation in the wavelet domain. The spurious outliers are modeled as the convolution of Dirac functions with the point spread function of the instrument (PSF). More precisely, the convolution kernel is modeled as a Gaussian function $\exp^{-\frac{\left((x-x_0)^2+(y-y_0)^2)\right)}{w}}$, where (x, y) denotes the position of the pixel , and (x_0, y_0), the pixel in the center of the image (the kernels are then normalized). In the following, the amplitude of $\mathbf{O}$ is fixed from one experiment to another (and so their energy increases with w). Consequently, this setting makes the particular choice of wavelet functions critical since it will largely impact the coherence between the $\mathbf{\Phi_S}$ and $\mathbf{\Phi_O}$. On the one hand, highly oscillating wavelet functions (*i.e.* with a large number of vanishing moments) will yield more incoherent dictionaries but at the cost of slightly less sparse representations for the sources. On the other hand, more localized wavelet functions are likely to provide better sparse representations but at the cost of lowering the morphological diversity between the dictionaries. Therefore, in this particular robust BSS problem, one needs to make a trade-off between the compressibility of the sparse representations, which is essential for source separation, and the morphological diversity between $\mathbf{\Phi_S}$ and $\mathbf{\Phi_O}$, which is of paramount importance for the separation of the sources and the outliers. In the next experiments, $\mathbf{\Phi_S}$ will be chosen as undecimated Daubechies wavelet transforms with varying vanishing moments.

6.3.2. Influence of the dictionary Φ_S. To further highlight the role played by the mutual coherence in the proposed tr-rGMCA algorithm, we propose to investigate the influence of the vanishing moments of the Daubechies wavelet functions used for $\mathbf{\Phi_S}$. We only compare the different methods that are influenced by the choice of $\mathbf{\Phi_S}$: the so-called oracle, tr-rGMCA, AMCA performed on $\mathbf{X}$ and $\mathbf{X}-\mathbf{O}$ (the combination MCA+AMCA performs so poorly that the influence of $\mathbf{\Phi_S}$ cannot be commented, and the influence of $\mathbf{\Phi_S}$ on OP+AMCA can be deduced by the performances of AMCA).

Vanishing Moments:	4	8	12	16	20
$\frac{\left\lVert \mathbf{S}\mathbf{\Phi}_\mathbf{S}^T \right\rVert_1}{\left\lVert \mathbf{S}\mathbf{\Phi}_\mathbf{S}^T \right\rVert_2}$	143.36	142.18	142.48	143.06	143.64
$\frac{\left\lVert H\mathbf{\Phi}_\mathbf{S}^T \right\rVert_1}{\left\lVert H\mathbf{\Phi}_\mathbf{S}^T \right\rVert_2}$	21.97	31.48	38.42	45.13	52.19

Table 2: Influence of the number of vanishing moments on the representation of $\mathbf{S}$ and the outliers.

First, the choice of $\mathbf{\Phi_S}$ does not significantly impact the AMCA algorithm that is performed by $\mathbf{X}-\mathbf{O}$ fig.12: the representation coefficients of the sources are sufficiently sparse, for the different dictionaries, to perform the unmixing (the ratio $\frac{\left\lVert \mathbf{S}\mathbf{\Phi}_\mathbf{S}^T \right\rVert_1}{\left\lVert \mathbf{S}\mathbf{\Phi}_\mathbf{S}^T \right\rVert_2}$, which somehow measures the level of sparsity of the sources in $\mathbf{\Phi_S}$, does not significantly change in table 2). However, one of the sources, the thermal dust emission is not very accurately recovered: it is

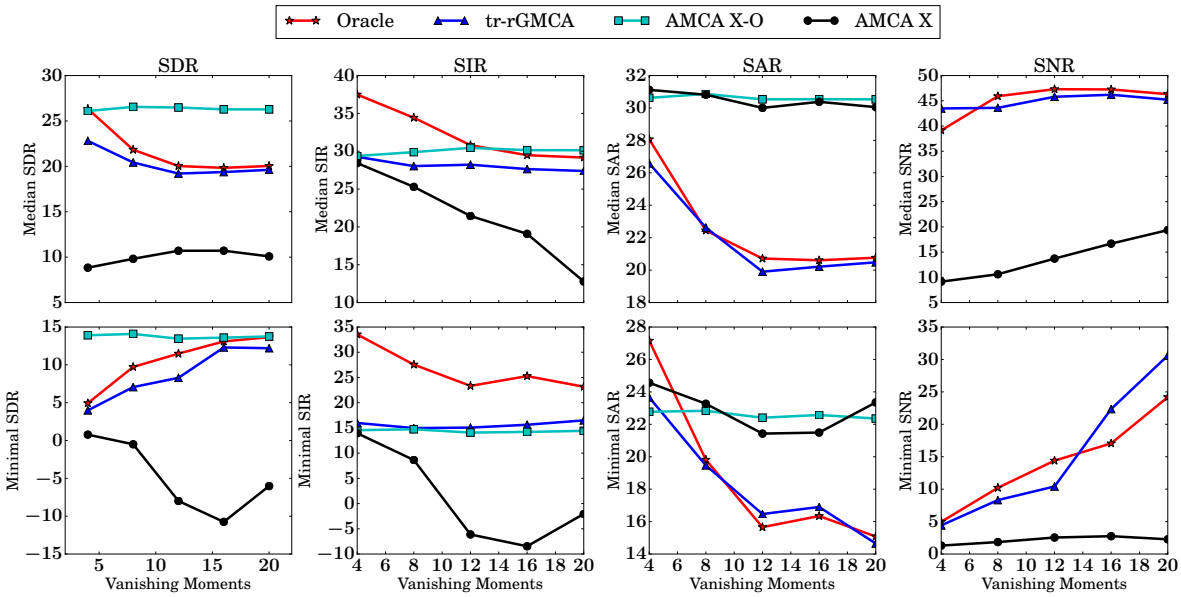


Figure 12: Performance indicators for a varying number of vanishing moments.

hampered by the correlations between sources and as well as their spectra, which are displayed in fig.11.

The influence of $\mathbf{\Phi_S}$ for the oracle and tr-rGMCA are similar. The SIR and SAR decrease when the number of vanishing moments increases. Indeed, the largest scales of the astrophysical sources are partially correlated. Therefore, the most discriminative coefficients in the wavelet domain are located in the finest wavelet scales, which is however the most coherent with $\mathbf{\Phi_O}$. This is especially true when the number of vanishing moments is low. On the other hand, the SNR values, especially the minimal SNR, increase: the outliers do not leak towards the estimated sources when the number of vanishing moment is large enough (the outliers are less sparsely represented in $\mathbf{\Phi_S}$, table 2).

In the following, we will make use of the Daubechies wavelets with 20 vanishing moments so as to recover all the sources fairly while providing an improved separation with respect to the outliers.

6.3.3. Influence of the kernel width. In this experiment, $\mathbf{\Phi_S}$ is fixed and the width of the Gaussian kernel w, which also tends to alter the coherence between $\mathbf{\Phi_S}$ and $\mathbf{\Phi_O}$ and as well the morphological diversity between $\mathbf{O}$ and $\mathbf{S}$, is varying. The kernel width w varies in Fig. 13.

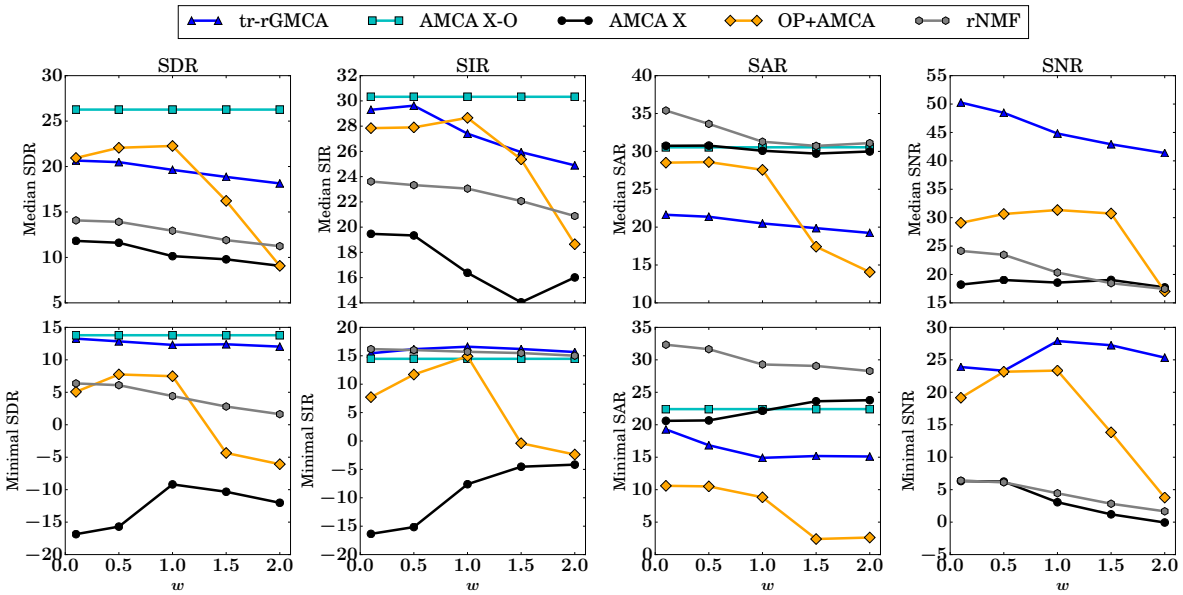


Figure 13: Performance indicators for a varying variance of the Gaussian kernel.

Variance - parameter w:	0.1	0.5	1	1.5	2
$\frac{\lVert \mathbf{S}\mathbf{\Phi}_{\mathbf{O}}^T \rVert_1}{\lVert \mathbf{S}\mathbf{\Phi}_{\mathbf{O}}^T \rVert_2}$ *	105.7	104.5	96.9	79.6	76.9

Table 3: Influence of the width of the Gaussian kernel on the representation of $\mathbf{S}$. * the sources are artificially deconvolved with $\mathbf{H}$.

As illustrated in fig.13, all the methods are impacted by the width of the kernel (*i.e.* the morphology of the outliers).
First, we recall that the unmixing is very difficult when w is small: the outliers contaminate the high frequency content of the sources, which is discriminant for the unmixing (the large scales of the sources are correlated). That is why the results are on the overall improved when w increases but is small.
On the other hand, we can notice that the methods are hindered by a large w. In that case, the "low-frequency "(similar to the kernel) content of the sources, which contains most of their energy, become highly sparse in $\mathbf{\Phi_O}$ tab.3: they leak towards the estimated outliers. Consequently, the SAR fig.13 decreases as w increases. This is especially true for the thermal dust emission (associated with the minimal SAR), whose singularities have a morphology very similar to the one of the kernel. The leakages are also reinforced by the fact that the large scales of the sources are correlated: the $\ell_{2,1}$ penalization in $\mathbf{\Phi_O}$ is less expensive than the ℓ_1 in $\mathbf{\Phi_S}$ for the correlations. Besides, the energy of the outliers on the coarse-scale of $\mathbf{\Phi_S}$ increases, but is not thresholded (because it is not sparse): this is clearly hampering the SNR when w is large.

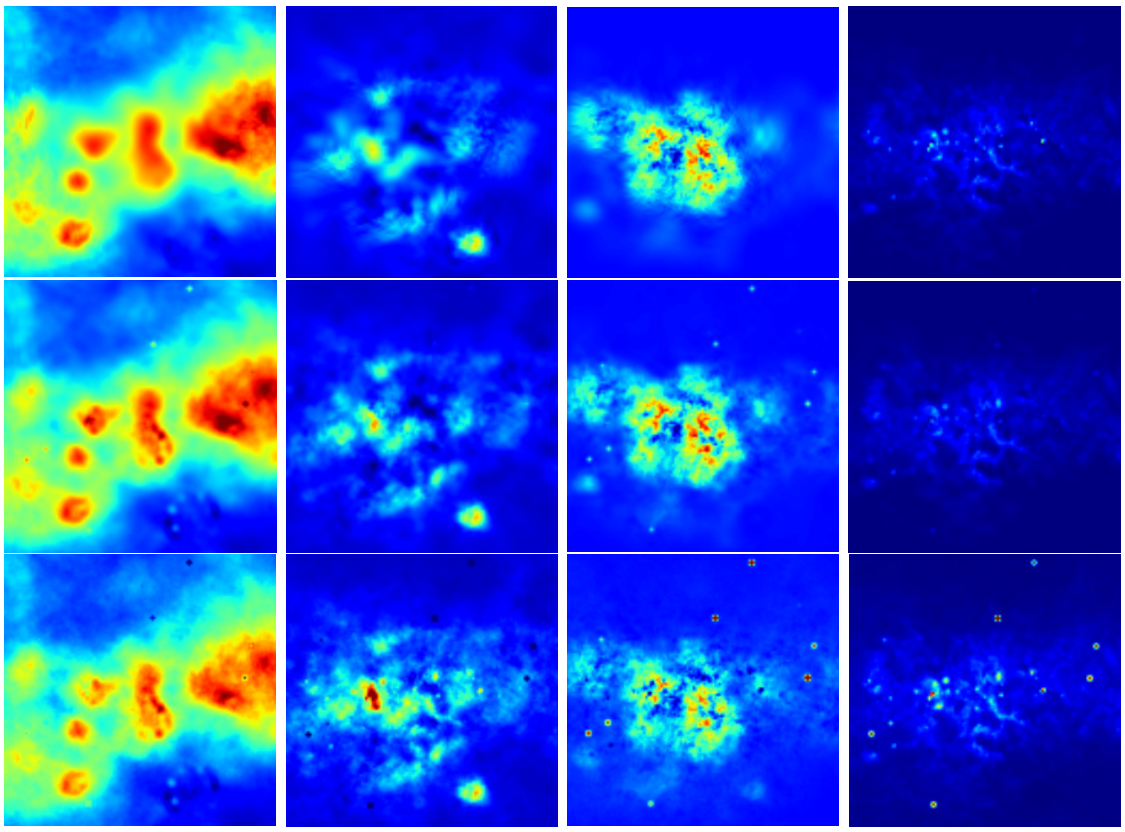


Figure 14: Estimated sources with tr-rGMCA (top row), OP+AMCA (second row) and rNMF (third row) with $w = 1$ and 20 vanishing moments for the wavelets.

Only the combination OP+AMCA is able to outperform tr-rGMCA in term of highest SDR, while the kernel is not too large. However, only tr-rGMCA is able to fairly recover the thermal dust emission, as well as AMCA performed on $\mathbf{X} - \mathbf{O}$. We underline that the parameter of OP was manually tuned knowing the ground truth, and there is no doubt that if the parameters involved in tr-rGMCA were similarly tuned, its performances would be, at least, similar to the ones of OP+AMCA. The rNMF method, even if it is initialized from the ground truth $\mathbf{A}$, was not able to correctly unmix the sources and separate the outliers from the source contribution: the fact that the sources samples do not lie in the simplex makes this method inefficient in this experiment since $\mathbf{O}$ cannot be separated from the $\mathbf{AS}$.
Illustrative results are provided in fig.14 and 15. Outlier residuals are present in the sources estimated by rNMF and OP+AMCA, fig.14. On the other hand, the highest frequency contributions of the sources is not correctly recovered by tr-rGMCA (the SAR are quite low fig.14), and have leaked towards the estimated outliers. The spectra recovered by tr-rGMCA are the most precise, in particular the other methods have fail to recover the thermal dust spectrum precisely fig.15.

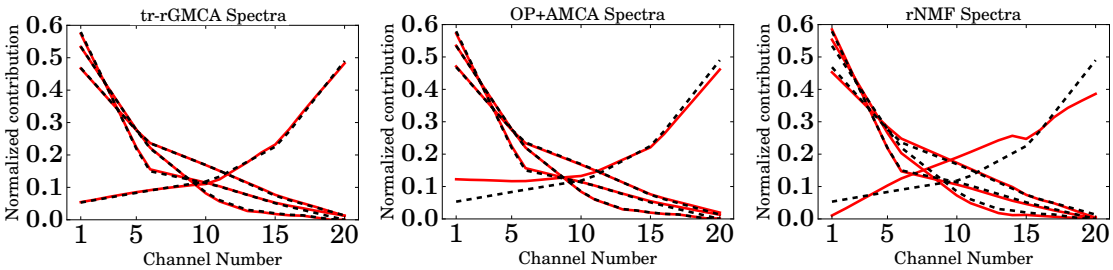


Figure 15: Estimated spectra, with $w = 1$ and 20 vanishing moments for the wavelets. Red lines: estimated spectra, black dashed lines: ground truth.

Code. Following the philosophy of reproducible research [10], a python implementation of the algorithms introduced in this article will be available at *https://www.cosmostat.org/software/gmcalab*.

7. Conclusion.

In this article, we introduced a new solution for the BSS problem in the presence of outliers that allows a robust estimation of the mixing matrix and an accurate separation of the sources and the outliers. The proposed tr-rGMCA algorithm estimates jointly the mixing matrix, the sources and the outliers so as to simultaneously unmix the sources and separate the outliers from the source contribution. Building upon sparse modeling, it first exploits the morpho-spectral diversity between the outliers and source contribution to distinguish between them, including in the challenging determined setting. The tr-rGMCA algorithm builds upon a two-stage optimization procedure: i) a warm-up stage based on heuristics that yield a reliable algorithm with enhanced robustness and ii) a refinement step based on the PALM algorithm that provably converges to a stationary point to the problem. Numerical experiments have been carried out on Monte-Carlo simulations which show the robustness of the proposed approach which provides state-of-the-art results. Future work will focus on extending the proposed approach to detect and estimate spectral variabilities in hyperspectral imaging.

Acknowledgments. This work is supported by the European Community through the grants PHySIS (contract no. 640174) and LENA (ERC StG no. 678282) within the H2020 Framework Program.

Appendix A. Proximal Operators.

Let $f : \mathbb{R}^{p\times q} \to \,]-\infty, +\infty]$, where $p, q \in \mathbb{N}$, be a proper, lower semi-continuous and convex function. Its proximal operator is given by $\mathrm{prox}_f : \mathbb{R}^{p\times q} \to \mathbb{R}^{p\times q}, x \mapsto \mathrm{argmin}_y \frac{1}{2}\|x - y\|_2^2 + f(y)$ [16].
In the following table, we present the different functions that are used in this article and their associated proximal operators.

Function	Proximal operator
$\chi_{\mathbf{Y}:\|\mathbf{Y}\|_2\leq 1}(\mathbf{X})$	$\mathbf{X}' : (\mathbf{X}')^i = \frac{\mathbf{X}^i}{\max(1,\|\mathbf{X}^i\|_2)} \forall i$ [16]
$\|\Lambda \odot \mathbf{X}\|_1$	$\mathcal{S}_\Lambda(\mathbf{X})$ [16]
$\left\|\Lambda \odot \mathbf{X}\mathbf{\Phi}_\mathbf{S}^T\right\|_1$	$\mathcal{S}_\Lambda\left(\mathbf{X}\mathbf{\Phi}_\mathbf{S}^T\right)\mathbf{\Phi}_\mathbf{S}$ [43] (exact if $\mathbf{\Phi}_\mathbf{S}$ is orthonormal and good approximation if transformation with diagonally dominant Gram matrix)
$\|\Upsilon \odot \mathbf{X}\|_{2,1}$	$\mathbf{X}' : (\mathbf{X}')^i = \mathbf{X}^i \times \left(1 - \frac{\Upsilon^i}{\|(\mathbf{X})^i\|_2}\right)_+, \forall i$, [28].
$\left\|\Upsilon \odot \mathbf{X}\mathbf{\Phi}_\mathbf{O}^T\right\|_{2,1}$	$\mathbf{X}'\mathbf{\Phi}_\mathbf{O} : (\mathbf{X}')^i = (\mathbf{X}\mathbf{\Phi}_\mathbf{O}^T)^i \times \left(1 - \frac{\Upsilon^i}{\left\|(\mathbf{X}\mathbf{\Phi}_\mathbf{O}^T)^i\right\|_2}\right)_+, \forall i$ (exact if $\mathbf{\Phi}_\mathbf{O}$ is orthonormal and good approximation if transformation with diagonally dominant Gram matrix)
$\|\Upsilon \odot \mathbf{X}\|_{2,1} + \chi_{\mathbf{Y}:\mathbf{Y}\geq 0}(\mathbf{X})$	$\mathbf{X}' : (\mathbf{X}')^i = \mathbf{X}_+^i \times \left(1 - \frac{\Upsilon^i}{\left\|(\mathbf{X})_+^i\right\|_2}\right)_+, \forall i$, [51, Theorem 1].

Similarly to $\left\|\Lambda.\mathbf{\Phi}_\mathbf{S}^T\right\|_1$, we do not find a closed form formulation for $\left\|\Upsilon.\mathbf{\Phi}_\mathbf{O}^T\right\|_{2,1}$ when $\mathbf{\Phi}_\mathbf{O}$ is not orthonormal. In the spirit of the approximation made for the ℓ_1 norm, we propose to threshold the columns of the expansion coefficients, and then come back to the domain of observations. In practice, these approximations made to handle sparsity in a transformed domain give better results than the synthesis formulation and supports the use of these approximations.

REFERENCES

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

[2] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.

[3] 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, 5 (2012), pp. 354–379.

[4] 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, 63 (2015), pp. 1199–1213, doi:10.1109/TSP.2015.2391071.

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

[6] J. Bobin, J.-L. Starck, J. Fadili, Y. Moudden, and D. Donoho, *Morphological component analysis: An adaptive thresholding strategy*, IEEE Trans. On Image Processing, 16 (2007), pp. 2675 – 2681.

[7] J. Bobin, F. Sureau, J.-L. Starck, A. Rassat, and P. Paykari, *Joint Planck and WMAP CMB map reconstruction*, A&A, 563 (2014).

[8] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494.

[9] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images*, SIAM Review, 51 (2009), pp. 34–81.

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

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

[12] E. J. Candes, M. B. Wakin, and S. P. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, Journal of Fourier analysis and applications, 14 (2008), pp. 877–905.

[13] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, *Rank-sparsity incoherence for matrix decomposition*, SIAM Journal on Optimization, 21 (2011), pp. 572–596, doi:10.1137/090761793.

[14] C. Chenot and J. Bobin, *Blind separation of sparse sources in the presence of outliers*, Signal Processing, 138 (2017), pp. 233 – 243.

[15] C. Chenot and J. Bobin, *Bss with corrupted data in transformed domains*, in Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings, 2017, pp. 542–552.

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

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

[18] J. Delabrouille and et al., *The pre-launch Planck Sky Model: a model of sky emission at submillimetre to centimetre wavelengths*, Astronomy & Astrophysics, 553 (2013), A96, p. A96, doi:10.1051/0004-6361/201220019, arXiv:1207.3675.

[19] D. L. Donoho, M. Elad, and V. N. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Transactions on information theory, 52 (2006), pp. 6–18.

[20] M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho, *Simultaneous cartoon and texture image inpainting using morphological component analysis (mca)*, Applied and Computational Harmonic Analysis, 19 (2005), pp. 340–358.

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

[22] 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.

[23] 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.

[24] N. Gillis and A. Kumar, *Exact and heuristic algorithms for semi-nonnegative matrix factorization*, SIAM Journal on Matrix Analysis and Applications, 36 (2015), pp. 1404–1424.

[25] Q. Ke and T. Kanade, *Robust ℓ_1 norm factorization in the presence of outliers and missing data by alternative convex programming*, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, IEEE, 2005, pp. 739–746.

[26] N. Keshava and J. F. Mustard, *Spectral unmixing*, IEEE signal processing magazine, 19 (2002), pp. 44–57.

[27] D. Kong, C. Ding, and H. Huang, *Robust nonnegative matrix factorization using $\ell_{2,1}$-norm*, in Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011, pp. 673–682.

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

[29] P. Kuppinger, G. Durisi, and H. Bolcskei, *Uncertainty relations and sparse signal recovery for pairs of general signal sets*, IEEE Transactions on Information Theory, 58 (2012), pp. 263–277.

[30] S. M. Leach, J.-F. Cardoso, C. Baccigalupi, R. Barreiro, M. Betoule, J. Bobin, A. Bonaldi, J. Delabrouille, G. De Zotti, C. Dickinson, et al., *Component separation methods for the planck mission*, Astronomy & Astrophysics, 491 (2008), pp. 597–615.

[31] 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, 7 (2014), pp. 3742–3754, doi:10.1109/JSTARS.2014.2360409.

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

[33] S. Nakhostin, H. Clenet, T. Corpetti, and N. Courty, *Joint anomaly detection and spectral unmixing for planetary hyperspectral images*, IEEE Transactions on Geoscience and Remote Sensing, 54

(2016), pp. 6879–6894.

[34] T.-H. Oh, Y.-W. Tai, J.-C. Bazin, H. Kim, and I. S. Kweon, *Partial sum minimization of singular values in robust pca: Algorithm and applications*, IEEE transactions on pattern analysis and machine intelligence, 38 (2016), pp. 744–758.

[35] P. Paatero and U. Tapper, *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics, 5 (1994), pp. 111–126.

[36] N. Parikh, S. P. Boyd, et al., *Proximal algorithms.*, Foundations and Trends in optimization, 1 (2014), pp. 127–239.

[37] H. Raguet, J. Fadili, and G. Peyré, *A generalized forward-backward splitting*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1199–1226.

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

[39] 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, (2015), p. 8.

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

[41] H.-J. M. Shi, S. Tu, Y. Xu, and W. Yin, *A primer on coordinate descent algorithms*, arXiv preprint arXiv:1610.00040, (2016).

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

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

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

[45] P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret, *Hyperspectral unmixing with spectral variability using a perturbed linear mixing model*, IEEE Transactions on Signal Processing, 64 (2016), pp. 525–538.

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

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

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

[49] H. Xu, C. Caramanis, and S. Sanghavi, *Robust pca via outlier pursuit*, in Advances in Neural Information Processing Systems, 2010, pp. 2496–2504.

[50] Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM Journal on imaging sciences, 6 (2013), pp. 1758–1789.

[51] Y.-L. Yu, *On decomposing the proximal map*, in Advances in Neural Information Processing Systems, 2013, pp. 91–99.

[52] A. Zare and K. Ho, *Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing*, IEEE Signal Processing Magazine, 31 (2014), pp. 95–104.

[53] 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, 52 (2014), pp. 4729–4743, doi:10.1109/TGRS.2013.2284280.

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

[55] T. Zhou and D. Tao, *Godec: Randomized low-rank & sparse matrix decomposition in noisy case*, in Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 33–40.

[56] Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma, *Stable principal component pursuit*, CoRR, abs/1001.2363 (2010), http://arxiv.org/abs/1001.2363.

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