Blind Source Separation: the Sparsity Revolution

J. Bobin, Jean-Luc Starck, Y. Moudden, Jalal M. Fadili

Blind Source Separation: the Sparsity Revolution

Bobin J. [a] Starck J.-L. [a] Moudden Y. [a] Fadili M.J. [b]

[a] *Laboratoire AIM, CEA/DSM-CNRS-Université Paris Diderot, CEA Saclay, IRFU/SEDI-SAP, Service d'Astrophysique, Orme des Merisiers, 91191 Gif-sur-Yvette, France.*

[b] *GREYC CNRS UMR 6072, Image Processing Group, ENSICAEN 14050, Caen Cedex, France.*

Abstract

Over the last few years, the development of multi-channel sensors motivated interest in methods for the coherent processing of multivariate data. Some specific issues have already been addressed as testified by the wide literature on the so-called blind source separation (BSS) problem. In this context, as clearly emphasized by previous work, it is fundamental that the sources to be retrieved present some quantitatively measurable diversity. Recently, sparsity and morphological diversity have emerged as a novel and effective source of diversity for BSS. We give here some essential insights into the use of sparsity in source separation and we outline the essential role of morphological diversity as being a source of diversity or contrast between the sources. This paper overviews a sparsity-based BSS method coined Generalized Morphological Component Analysis (GMCA) that takes advantages of both morphological diversity and sparsity, using recent sparse overcomplete or redundant signal representations. GMCA is a fast and efficient blind source separation method. In remote sensing applications, the specificity of hyperspectral data should be accounted for. We extend the proposed GMCA framework to deal with hyperspectral data. In a general framework, GMCA provides a basis for multivariate data analysis in the scope of a wide range of classical multivariate data restorate. Numerical results are given in color image denoising and inpainting. Finally, GMCA is applied to the simulated ESA/Planck data. It is shown to give effective astrophysical component separation.

Key words: Blind component separation, Sparse overcomplete representations, Sparsity, Morphological component analysis, Morphological diversity

1 Introduction

Finding a suitable representation of multivariate data is a longstanding problem in statistics and related areas. Good representation means that the data is somehow transformed so that its essential structure is made more visible or more easily accessible. This problem is for instance encountered in unsupervised learning, exploratory data analysis and signal processing. In the latter, a typical field where the good representation problem arises is source separation. Over the last few years, the development of multi-channel sensors motivated interest in such methods for the coherent processing of multivariate data. Areas of application include biomedical engineering, medical imaging, speech processing, astronomical imaging, remote sensing, communication systems, seismology, geophysics, econometrics.

Consider a situation where there is a collection of signals emitted by some physical objects or sources. These physical sources could be, for example, different brain areas emitting electric signals; people speaking in the same room (the classical cocktail party problem), thus emitting speech signals; or radiation sources emitting their electromagnetic waves. Assume further that there are several sensors or receivers. These sensors are in different positions, so that each records a mixture of the original source signals with different weights. It is assumed that the mixing weights are unknown, since their knowledge entails knowing all the properties of the physical mixing system, which is not accessible in general. Of course, the source signals are unknown as well, since the primary problem is that they cannot be recorded directly. The *blind source separation* (BSS) problem is to find the original signals from their observed mixtures, without prior knowledge of the mixing weights, and by knowing very little about the original sources. In the classical example of the cocktail party, the BSS problem amounts to recovering the voices of the different speakers, from the mixtures recorded at several microphones.

There has been a flurry of research activity on BSS which is one of the hottest areas in the signal processing community. Some specific issues have already been addressed using a blend of heuristic ideas and rigorous derivations. This is testified by the extensive literature on the subject. As clearly emphasized by previous work, it is fundamental that the sources to be retrieved present some quantitatively measurable diversity (*e.g.* decorrelation, independence, morphological diversity, etc). Recently, sparsity and morphological diversity have emerged as a novel and effective source of diversity for BSS.

The goal of this paper is to give some new and essential insights into the use of sparsity in source separation and to outline the fundamental role of morphological diversity as being a source of diversity or contrast between the sources. This paper describes a BSS method, and more generally a multichan-

nel sparsity-based data analysis framework, coined Generalized Morphological Component Analysis (GMCA) which is fast, efficient and robust to noise. GMCA takes advantages of both morphological diversity and sparsity, using recent sparse overcomplete signal representations. Theoretical arguments as well as numerical experiments in multivariate image processing are reported to characterize and illustrate the good performance of GMCA for BSS.

1.1 Organization of the paper

In Section 2 we formally state the BSS problem and survey the current state-of-the-art in the field of BSS. In Section 3, we give the necessary background on sparse overcomplete representation and decomposition, with extensions to the multichannel setting. In Section 4, the multichannel extension of the Morphological Component Analysis (MCA) algorithm is described and some of its theoretical properties are stated. In Section 5, a new way of thinking sparsity in BSS. All necessary ingredients introduced in previous sections are put together, and the GMCA algorithm for BSS is provided. The extension of GMCA to hyperspectral data and its application of GMCA to multichannel data restoration analysis are reported in 6 and Section 7.1. We also discuss an application of the GMCA BSS algorithm to an astronomical imaging experiment.

1.2 Definitions and notations

Unless stated otherwise, a vector x will be a row vector $x = [x_1, \cdots, x_t]$. We equip the vector space $\mathbb{R}^t$ with the scalar product $\left\langle x, y \right\rangle = xy^T$. The ℓ_p-norm of a vector x is defined by $\|x\|_{\ell_p} = (\sum_i |x[i]|^p)^{1/p}$, with the usual notation $\|x\|_{\ell_\infty} = \max_i |x[i]|$. The notation $\|x\|_{\ell_0}$ defines the ℓ_0 quasi-norm of x (*i.e.* the number of non-zero elements in x).

Bold symbols represent matrices and $\mathbf{X}^T$ is the transpose of $\mathbf{X}$. The i-th entry of x_p is $x_p[i]$, x_p is the p-th row and x^q the q-th column of $\mathbf{X}$. The "entrywise" p-norm of a matrix $\mathbf{X}$ is defined by $\|\mathbf{X}\|_p = \left(\sum_{i,j} |x_i[j]|^p\right)^{1/p}$, not to be confused with matrix induced p-norms. The Frobenius norm of $\mathbf{X}$ is obtained for $p = 2$, $\|\mathbf{X}\|_F^2 = \text{Trace}\left(\mathbf{X}^T\mathbf{X}\right)$. Similarly to vectors, $\|\mathbf{X}\|_\infty$ and $\|\mathbf{X}\|_0$ respectively denote the maximum in magnitude and the number of nonzero entries in the matrix $\mathbf{X}$.

In the proposed iterative algorithms, $\tilde{x}^{(h)}$ will be the estimate of x at iteration h. $\mathbf{\Phi} = [\phi_1^T, \cdots, \phi_T^T]^T$ defines a $T \times t$ dictionary the rows of which are unit ℓ_2-norm atoms $\{\phi_i\}_i$. The mutual coherence of $\mathbf{\Phi}$ (see (1) and references

therein) is $\mu_{\mathbf{\Phi}} = \max_{i \neq j} \left| \left\langle \phi_i, \phi_j \right\rangle \right|$. When $T > t$, this dictionary is said to be redundant or overcomplete. In the next, we will be interested in the decomposition of a signal x in $\mathbf{\Phi}$. We thus define $\mathcal{S}^{\mathbf{\Phi}}_{\ell_0}(x)$ (respectively $\mathcal{S}^{\mathbf{\Phi}}_{\ell_1}(x)$) the set of solutions to the minimization problem $\min_c \|c\|_{\ell_0}$ s.t. $x = c\mathbf{\Phi}$ (respectively $\min_c \|c\|_{\ell_1}$ s.t. $x = c\mathbf{\Phi}$). When the ℓ_0 sparse decomposition of a given signal x has a unique solution, let $\alpha = \Delta_{\mathbf{\Phi}}(x)$ where $x = \alpha\mathbf{\Phi}$ denote this solution. Finally, we define $\Delta_\lambda(.)$ to be a thresholding operator with threshold λ (hard-thresholding or soft-thresholding; this will be specified when needed).

The support $\Lambda(x)$ of row vector x is $\Lambda(x) = \{k; |x[k]| > 0\}$. Note that the notion of support is well-adapted to ℓ_0-sparse signals as these are synthesized from a few non-zero dictionary elements. Similarly, we define the δ-support of x as $\Lambda_\delta(x) = \{k; |x[k]| > \delta \|x\|_{\ell_\infty}\}$.

2 Blind Source Separation, a strenuous inverse problem

2.1 Modelling multichannel data

2.1.1 The BSS model

In a source separation setting, the observed data are composed of m distinct *monochannel* datum $\{x_i\}_{i=1,\cdots,m}$. Each datum could be a $\sqrt{t} \times \sqrt{t}$ image or a monodimensional signal with t samples. In the next, we assume that each observation $\{x_i\}_{i=1,\cdots,m}$ is a row-vector of size t. The classical instantaneous linear mixture model states that each datum is the linear combination of n so-called sources $\{s_j\}_{j=1,\cdots,n}$ such that:

$$\forall i = 1, \cdots, m; \quad x_i = \sum_{j=1}^{n} a_{ij} s_j \ , \tag{1}$$

where the set of scalar values $\{a_{ij}\}_{i=1,\cdots,m;j=1,\cdots,n}$ models the "weight" of each source in the composition of each datum. For convenience, the mixing model with additive noise can be rewritten in matrix form:

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

where $\mathbf{X}$ is the $m \times t$ measurement matrix, $\mathbf{S}$ is the $n \times t$ source matrix and $\mathbf{A}$ is the $m \times n$ mixing matrix. $\mathbf{A}$ defines the contribution of each source to each measurement. An $m \times t$ matrix $\mathbf{N}$ is added to account for instrumental noise or model imperfections. In this paper, we will only study the overdetermined case: $m \geq n$; the converse underdetermined case ($m < n$) is a more difficult problem (see (2; 3) for further details). Further work will be devoted to this particular case.

In the blind source separation problem, both the mixing matrix $\mathbf{A}$ and the sources $\mathbf{S}$ are unknown and must be estimated jointly. In general, without further *a priori* knowledge, decomposing a rectangular matrix $\mathbf{X}$ into a linear combination of n rank-1 matrices is clearly ill-posed. The goal of BSS is to understand the different cases in which this or that additional prior constraint allows to reach the land of well-posed inverse problems and to devise separation methods that can handle the resulting models.

2.1.2 A question of diversity

Note that the mixture model in Equation (1) is equivalent to the following one:

$$\mathbf{X} = \sum_{i=1}^{n} a^i s_i \ , \tag{3}$$

where a^i is the i-th column of $\mathbf{A}$. Blind Source Separation is equivalent to decomposing the data $\mathbf{X}$ into a sum of n rank-1 matrices $\{\mathbf{X}_i = a^i s_i\}_{i=1,\cdots,n}$. Obviously, there are infinitely many ways of decomposing a given matrix with rank n into the linear combination of n rank-1 matrices. Further information is required to disentangle between the sources.
Let us assume that the sources are random vectors. These may be known *a priori* to be different in the sense of being simply decorrelated. A separation scheme will then look for sources $\mathbf{S}$ such that their covariance matrix $\mathbf{R_S}$ is diagonal. Unfortunately, the covariance matrix $\mathbf{R_S}$ is invariant by orthonormal transformations such as rotations.

Therefore, an effective BSS method has to go beyond decorrelation (see (4; 5) for further reflections about the need for stronger *a priori* constraints going beyond the decorrelation assumption).
In the next sections we will emphasize on different sets of *a priori* constraints and different methods to handle them. In Section 2.2, we give an overview of BSS methods that use statistical independence as the key assumption for separation. Recently, sparsity has emerged as being a very effective way to distinguish the sources. These new approaches are introduced in Section 2.3.

2.2 Independent Component Analysis

2.2.1 Generalities

The previous section emphasized on the need for further "a priori" assumptions to bring blind source separation to the "land" of well-posed inverse problems. In this section, we cope with noiseless mixtures assuming that $\mathbf{X} = \mathbf{AS}$. The case where the data are perturbed by additive noise will be discussed at

the end of this section.
The seminal work by Comon (6) paved the way for the outgrowth of *Independent Component Analysis* (ICA). In the celebrated ICA framework, the sources are assumed to be independent random variables with joint probability density function $f_{\mathbf{S}}$ such that:

$$f_{\mathbf{S}}(s_1, \cdots, s_n) = \prod_{i=1}^{n} f_{s_i}(s_i) \ . \tag{4}$$

Disentangling between sources requires a way to measure how separable sources are different. As statistical independence is verified by the *pdf* of the sources, devising a good "measure" of independence is not trivial. In that setting, ICA then boils down to finding a multichannel representation/basis on which the estimated sources $\tilde{\mathbf{S}}$ are as "independent as possible". Equivalently, ICA looks for a *separating/demixing* matrix $\mathbf{B}$ such that the estimated sources $\tilde{\mathbf{S}} = \mathbf{BAS}$ are independent. Until the end of the section devoted to ICA, we will assume that the mixing matrix $\mathbf{A}$ is a square invertible matrix ($m = n$ and $\det(\mathbf{A}) > 0$).
Until now, we can wonder if independence makes the sources identifiable. Under mild conditions, the Darmois theorem (7) shows that statistical independence means separability (6). It states that if at most one of the sources is generated from a Gaussian distribution then if the entries of $\tilde{\mathbf{S}} = \mathbf{BAS}$ are independent then $\mathbf{B}$ is a separating matrix and $\tilde{\mathbf{S}}$ is equal to $\mathbf{S}$ up to a scale factor (multiplication by a diagonal matrix with strictly positive diagonal entries) and permutation. As a consequence, if at most one source is Gaussian, maximizing independence between the estimated sources leads to perfect estimation of $\mathbf{S}$ and $\mathbf{A} = \mathbf{B}^{-1}$. The Darmois theorem then motivates the use of independence in blind source separation. It paved the way for the popular Independent Component Analysis (ICA).

2.2.1.1 Independence and Gaussianity : The Kullback-Leibler (KL) divergence from the joint density $f_{\mathbf{S}}(s_1, \cdots, s_n)$ to the product of its marginal density is a popular measure of statistical independence :

$$\mathcal{J}(\mathbf{S}) = \mathcal{K}\left[f_{\mathbf{S}}(s_1, \cdots, s_n), \prod_{i=1}^{n} f_{\mathbf{S}}(s_i)\right] \tag{5}$$

$$= \int_{\mathbf{S}} f_{\mathbf{S}}(s_1, \cdots, s_n) \log\left(\frac{f_{\mathbf{S}}(s_1, \cdots, s_n)}{\prod_{i=1}^{n} f_{\mathbf{S}}(s_i)}\right) \ , \tag{6}$$

Interestingly (see (8)), the KL can be decomposed into two terms as follows:

$$\mathcal{J}(\mathbf{S}) = \mathcal{C}(\mathbf{S}) - \sum_{i=1}^{n} \mathcal{G}(s_i) + K \ , \tag{7}$$

where $\mathcal{C}(\mathbf{S}) = \mathcal{K}[\mathcal{N}(\mathbb{E}\{\mathbf{S}\}, \mathbf{R_S}), \mathcal{N}(\mathbb{E}\{\mathbf{S}\}, \mathrm{diag}(\mathbf{R_S}))]$ and $\mathcal{G}(s_i) = \mathcal{K}\left[f(s_i), \mathcal{N}\left(\mathbb{E}\{s_i\}, \sigma^2_{s_i}\right)\right]$, $\sigma^2_{s_i}$ is the variance of s_i, and $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ is the normal probability density function with mean $\mathbf{m}$ and covariance $\mathbf{\Sigma}$. In Equation (7) K is a constant. The first term in Equation (7) vanishes when the sources are decorrelated. The second term measures the marginal **Gaussianity** of the sources. This decomposition of the KL entails that maximizing independence is equivalent to minimizing the correlation between the sources and maximizing their non-Gaussianity. Note that, with a taste of the central limit theorem, intuition tells us that mixing independent signals should lead to a kind of Gaussianization. It then seems natural that demixing leads to processes that deviate from Gaussian processes.

2.2.2 The algorithmic viewpoint

2.2.2.1 Approximating independence : In the ICA setting, the mixing matrix is square and invertible. Solving a BSS problem is equivalent to looking for a demixing matrix $\mathbf{B}$ that maximizes the independence of the estimated sources: $\tilde{\mathbf{S}} = \mathbf{BX}$. In that setting maximizing the independence of the sources (with respect to the Kullback-Leibler divergence) is equivalent to maximizing the non-Gaussianity of the sources. Since the seminal paper of Comon (6), a variety of ICA algorithms have been proposed. They all merely differ in the way they devise assessable *quantitative measures of independence*. Some popular approaches have given "measures" of independence are presented below :

- Information Maximization : (see (9; 10)) Bell and Sejnowski showed that maximizing the information of the sources is equivalent to minimizing the measure of independence based on the Kullback-Leibler divergence in Equation (5).

- Maximum Likelihood : Maximum Likelihood has also been proposed to solve the BSS issue. The Maximum Likelihood (ML) approach ((11; 12; 13)) has been showed to be equivalent to information maximization (InfoMax) in the ICA framework.

- Higher Order Statistics : As we pointed out earlier, maximizing the independence of the sources is equivalent to maximizing their non-Gaussianity under a strict decorrelation constraint. Because Gaussian random variables have vanishing higher order cumulants, devising a separation algorithm based on higher order cumulants should provide a way of accounting for the non-Gaussianity of the sources. A wide range of algorithms have been proposed based on the use of higher order statistics ((14; 15; 16), and references therein). Historical papers (see (6)) proposed ICA algorithms that

use approximations of the Kullback-Leibler divergence (based on truncated Edgeworth expansions). Interestingly, those approximations explicitly involve higher order statistics.

Lee et al. (see (17)) showed that most ICA-based algorithms are similar in theory and in practice.

2.2.2.2 Limits of ICA : Despite its theoretical strength and elegance, ICA suffers from several limitations:

- Probability density assumption : Even implicit, ICA algorithm requires information on the sources distribution. As stated in (17), whatever the contrast function to minimize (mutual information, ML, higher order statistics), most ICA algorithms can be equivalently restated in a *natural gradient form* ((18; 19)). In such setting, the "demixing" matrix $\mathbf{B}$ is estimated iteratively: $\mathbf{B} \leftarrow \mathbf{B} + \mu\Delta\mathbf{B}$ where the natural gradient of $\mathbf{B}$ is given by:

$$\Delta\mathbf{B} \propto \left[\mathbf{I} - h(\tilde{\mathbf{S}})\tilde{\mathbf{S}}^T\right]\mathbf{B} \ , \tag{8}$$

  where the function h is applied elementwise: $h(\tilde{\mathbf{S}}) = [h(\tilde{s}_{ij})]$ and $\tilde{\mathbf{S}}$ is the current estimate of $\mathbf{S}$: $\tilde{\mathbf{S}} = \mathbf{BX}$. Interestingly, the so-called *score* function h in Equation (8) is closely related to the assumed *pdf* of the sources (see (20; 19)). Assuming that all the sources are generated from the same probability density function $f_{\mathbf{S}}$, the so-called *score* function h is defined as follows:

$$h(\tilde{\mathbf{S}}) = -\frac{\partial \log\left(f_{\mathbf{S}}(\tilde{\mathbf{S}})\right)}{\partial\tilde{\mathbf{S}}} \ . \tag{9}$$

  As expected, the way the "demixing" matrix (and thus the sources) is estimated closely depends on the way the sources are modeled (from a statistical point of view). For instance, separating platykurtic (distribution with negative kurtosis) or leptokurtic (distribution with positive kurtosis) sources will require completely different score functions. Even if ICA is shown in (19) to be quite robust to "mis-modeling", the choice of the score function is crucial with respect to the convergence (and rate of convergence) of ICA algorithms. Some ICA-based techniques (see (21)) emphasized on adapting the popular FastICA algorithm to adjust the score function to the distribution of the sources. They particularly emphasize on modeling sources the distribution of which belongs to specific parametric classes of distributions such as generalized Gaussian: $f_{\mathbf{S}}(\mathbf{S}) \propto \prod_{ij} \exp(-\mu|s_{ij}|^{\theta})$ [1] .

[1] Note that the class of generalized Gaussian contains well-known distributions: the Gaussian ($\theta = 2$) and the Laplacian ($\theta = 1$) distributions.

- Noisy ICA : Only a few works have already investigated the problem of noisy ICA (see (22; 23)). As pointed out by Davies in (22), noise clearly degenerates the ICA model: it is not fully identifiable. In the case of additive Gaussian noise as stated in Equation (2), using higher order statistics yields an efficient estimate of the mixing matrix $\mathbf{A} = \mathbf{B}^{-1}$ (higher order statistics are blind to additive Gaussian noise; this property does not hold for non-Gaussian noise). Further, in the noisy ICA setting, applying the demixing matrix to the data does not yield an efficient estimate of the sources. Furthermore, most ICA algorithms assume the mixing matrix $\mathbf{A}$ to be square. When there is more observations than sources ($m > n$), a dimension reduction step is pre-processed. When noise perturbs the data, this subspace projection step can dramatically deteriorate the performance of the separation stage.

In the next Section we will introduce a new way of modeling the data so as to avoid most of the aforementioned limitations of ICA.

2.3 Sparsity in Blind Source Separation

In the above paragraph, we pointed out that Blind Source Separation is overwhelmingly a question of contrast and diversity. Indeed, devising a source separation technique consists in finding an effective way of disentangling between the sources. From this viewpoint, statistical independence is a kind of "measure" of diversity between signals. Within this paradigm, we can wonder if independence is a *natural* way of differentiating between signals.
As a statistical property, independence is a non-sense in a non-asymptotic study. In practice, one has to deal with finite-length signals; sometimes with a few samples. Furthermore, most real-world data are badly modeled by stationary stochastic processes. Let us consider the images in Figure 1.

Natural pictures are clearly non-stationary. As these pictures are slightly correlated, independence will fail in differentiating between them. Hopefully, the human eye (more precisely the different levels of the human visual cortex) is able to distinguish between those two images. Then, *what makes the eye so effective in discerning between visual "signals" ?*
The answer may come from neurosciences. Indeed, for a decades, many researchers (Barlow (24), Hubel and Wiesel [2], Olshausen (25), Field (26), Simoncelli (27) and references therein) in this field have endeavored to provide some exciting answers: the mammalian visual cortex seems to have learned via the natural selection of individuals, an effective way of coding the information in natural scenes. Indeed, the first level of the mammalian visual cortex

[2] Hubel and Wiesel were awarded with the Nobel Prize in Medicine in 1981.

Fig. 1. Examples of natural images.

(coined V1) seems to verify several interesting properties: i) it tends to "decorrelate" the responses of visual receptive fields (following Simoncelli *et al* (27), an efficient coding cannot duplicate information in more than one neuron), ii) owing to a kind of "economy/compression principle", saving neurons' activity yields a sparse activation of neurons for a given stimulus (this property can be considered as a way of compressing information).
Furthermore, the primary visual cortex is sensitive to particular stimuli (visual features) that surprisingly look like oriented Gabor-like wavelets (see (26)). It gives support to the crucial part played by contours in natural scenes. Furthermore, each stimulus tends to be coded by a few neurons. Such a way of coding information is often referred to as *Sparse Coding*. These few elements of neuroscience motivate the use of sparsity as an effective way of *compressing* signal's information thus extracting its very essence.
Inspired by the behavior of our visual cortex, seeking a sparse code may provide an effective way of differentiating between "different" signals. Here, "different" signals are signals with different sparse representations.

A pioneering work in sparse BSS

The seminal paper of Zibulevsky and Pearlmutter (28) introduced sparsity as an alternative to standard contrast functions in ICA. In this paper, the authors proposed to estimate the mixing matrix $\mathbf{A}$ and the sources $\mathbf{S}$ in a fully Bayesian framework. Each source $\{s_i\}_{i=1,\cdots,n}$ is assumed to be sparsely represented in the basis $\mathbf{\Phi}$:

$$\forall i = 1, \cdots, n; \quad s_i = \sum_{k=1}^{t} \alpha_i[k]\phi_k \ . \tag{10}$$

As the sources are assumed to be sparse, the distribution of their coefficients in Φ is a "sparse" (*i.e.* leptokurtic) prior distribution:

$$f_{\mathbf{S}}(\alpha_i[k]) \propto e^{-\mu_i g_\gamma(\alpha_i[k])} \ , \tag{11}$$

where $g_\gamma(\alpha_i[k]) = |\alpha_i[k]|^\gamma$ with $\gamma \leq 1$ [3]. Zibulevsky proposed to estimate $\mathbf{A}$ and $\mathbf{S}$ via a Maximum A Posteriori (MAP) estimator. The optimization task is then run using a Newton-like algorithm: the Relative Newton Algorithm (RNA - see (29) for more details). This new sparsity-based method paved the way for the use of sparsity in Blind Source Separation. Note that several other works emphasized the use of sparsity in a parametric Bayesian approach ((30) and references therein). Recently, sparsity has emerged as being an effective tool for solving underdetermined source separation issues ((31; 3; 32; 33) and references therein). In this paper, we will concentrate on overdetermined Blind Source Separation ($m \geq n$). Inspired by the work of Zibulevsky, we present a novel sparsity-based source separation framework providing new insights into BSS.

3 Sparse multichannel signal representation

3.1 The blessing of sparsity and overcomplete signal representations

In the last section we emphasized on the crucial role played by sparsity in BSS. Indeed, sparse representations provide an effective way to "compress" signals to a few very significant content. In previous work (see (34; 35)), we claimed that *the sparser the signals are, the better the separation is*. Therefore, the first step towards separation consists in finding an effective sparse representation; where "effective" means very sparse. Owing to its essential role in BSS, this section particularly emphasizes on the quest for sparse representation.

What's at stake : In the last decade sparsity has emerged as one of the leading concepts in a wide range of signal processing applications (restoration (36), feature extraction (37), source separation (38; 28; 39), compression ((40)), to name only a few). Sparsity has long been a theoretical and practical attractive signal property in many areas of applied mathematics (Computational harmonic analysis ((41)), Statistical estimation (42; 43)).
Very recently, researchers have advocated the use of overcomplete signal representations. Indeed, the attractiveness of redundant signal representations

[3] Applying $g_\gamma(.)$ pointwisely to a vector α_i is equivalent to computing its ℓ_γ norm.

relies on their ability to sparsely represent a large class of signals. Furthermore, handling very sparse signal representations allows more flexibility and entails effectiveness in many signal processing tasks (restoration, separation, compression, estimation, etc). Neuroscience also underlined the role of overcompleteness. Indeed, the mammalian visual system has been shown to be probably in need of overcomplete representation (25). In that setting, overcomplete *Sparse Coding* may lead to more effective (sparser) codes. In signal processing, both theoretical and practical arguments (44; 45) have supported the use of overcompleteness. It entails more flexibility in representation and effectiveness in many image processing tasks.
In the general sparse representation framework, a line vector signal $x \in \mathbb{R}^t$ is modeled as the linear combination of T elementary waveforms (the so-called *signal atoms*):

$$\{\phi_i\}_{i=1,\cdots,T}; \quad x = \sum_{i=k}^{T} \alpha[k]\phi_k \ , \tag{12}$$

where $\alpha[k] = \left\langle x, \phi_k \right\rangle$ are called the decomposition coefficients of x in the dictionary $\mathbf{\Phi} = [\phi_1^T, \cdots, \phi_T^T]^T$ (the $T \times t$ matrix whose rows are the atoms normalized to a unit ℓ_2-norm). In the case of overcomplete representations, the number of waveforms $\{\phi_k\}$ that composes the dictionary $\mathbf{\Phi}$ is higher than the dimension of the space in which x lies: $T > t$. In practice, the dimensionality of the sparse decomposition (*i.e.* the vector of coefficients α) can be very high: $T \gg t$.
Nonetheless, handling overcomplete representations is clearly an ill-posed problem owing to elementary linear algebra. Indeed *decomposing a signal in an overcomplete representation requires solving an underdetermined linear problem with more unknowns than data : $T > t$*. Linear algebra tells us that the problem $x = \alpha\mathbf{\Phi}$ has no unique solution. The next Section will provide solutions to this puzzling issue.

3.2 The sparse decomposition issue

In the sparse decomposition framework, the transition from ill-posedness to well-posedness is often fulfilled by reducing the space of candidate solutions to those satisfying some side constraints. Researchers have emphasized on adding a sparsity constraint to the previous ill-posed problem. Amongst all the solutions of $x = \alpha\mathbf{\Phi}$ we would like the sparsest one (with the least number of non-zero coefficients α_i). Donoho and Huo (46) proposed to solve the following minimization problem :

$$\min_{\alpha} \|\alpha\|_{\ell_0} \text{ s.t } x = \alpha\mathbf{\Phi} \ . \tag{13}$$

Clearly this is a *combinatorial* optimization problem that requires enumerating all the combinations of atoms $\{\phi_i\}_{i=1,\cdots,T}$ that synthesize x. This NP-hard

problem then appears hopeless. Donoho and Huo (46) proposed to relax the non-convex ℓ_0 sparsity by substituting the problem in Equation (13) with the following convex problem :

$$\min_{\alpha} \|\alpha\|_{\ell_1} \text{ s.t. } x = \alpha\mathbf{\Phi} \ . \tag{14}$$

The problem in Equation (14) is known as *Basis Pursuit* (see (47)). However, the solutions to the ℓ_0 and ℓ_1 problems are not equivalent in general. An extensive work (46; 48; 49; 50; 1; 51; 52) has focused on conditions under which the problems in Equation (13) and (14) are equivalent. Consider that $x = \sum_{k\in\Lambda(x)} \alpha[k]\phi_k$, we recall that $\Lambda(x)$ is the support of x in $\mathbf{\Phi}$ and $K = \text{Card}\left(\Lambda(x)\right)$. The signal x is said to be K-sparse in $\mathbf{\Phi}$. Interestingly, the first seminal work addressing the uniqueness and equivalence of the solutions to the ℓ_0 and ℓ_1 sparse decomposition recovery emphasized essentially on the structure of the overcomplete dictionary $\mathbf{\Phi}$. One quantitative measure that gives information about the structure of an overcomplete dictionary is its *mutual coherence* $\mu_{\mathbf{\Phi}}$, see also 1.2:

$$\mu_{\mathbf{\Phi}} = \max_{i\neq j} \left|\left\langle \phi_i, \phi_j \right\rangle\right| \ . \tag{15}$$

This parameter can be viewed as a worst-case measure of resemblance between all pairs of atoms. Interestingly, (46) showed that if a vector x^* with $\text{Card}\left(\Lambda(x^*)\right) = K$ is sufficiently sparse and verifies:

$$K < \frac{1}{2}\left(1 + \frac{1}{\mu_{\mathbf{\Phi}}}\right) \ , \tag{16}$$

then x^* is the unique maximally sparse solution to the ℓ_0 sparse decomposition problem in Equation (13), and the ℓ_0 and ℓ_1 sparse decomposition problems are equivalent. Consequently, recovering sparse decompositions is then made tractable. Note however that despite its simplicity, the identifiability test of (16) is pessimistic (worst-case analysis). More involved, but sharper, bounds of identifiability and equivalence between ℓ_0 and ℓ_1 problems have been proposed in the literature, see *e.g.* (49; 53; 52; 1; 48) and (54) for an extensive review.

3.3 Overcomplete multichannel representations

In this section we extend the sparse decomposition problem to the multichannel case. Previous work on the subject includes (55; 56) where all channels are constrained to have a common sparsity pattern (*i.e.* joint support), (57) in which the sparsity measure they used is different thus leading to different constraints, (58) which introduced the concept of multichannel dictionary. In this paper, we address a more general problem as we assume no constraint on the sparsity pattern of the different channels. Extending the redundant

representation framework to the multichannel case requires defining *what a multichannel overcomplete representation is*. We assume that the *multichannel* dictionary $\mathbf{\Psi}$ at hand is the tensor product of a *spectral* dictionary $\mathbf{\Xi}$ ($m \times n$ matrix) and a *spatial* or *temporal* dictionary $\mathbf{\Phi}$ ($T \times t$ matrix) [4]. Each atom of $\mathbf{\Psi}$ is then the tensor product of an atomic spectrum ξ_i and a spatial elementary signal ϕ_j:

$$\forall \{i,j\} \in \{1,\cdots,n\} \times \{1,\cdots,T\}, \quad \psi_{ij} = \xi_i \otimes \phi_j \ . \tag{17}$$

Recall that most popular sparse recovery results in the monochannel setting rely on the *mutual coherence* of the dictionary. In the multichannel case a similar quantity can be defined. Recalling the definition of mutual coherence given in section 1.2, the mutual coherence for multichannel dictionaries is as follows:

$$0 \leq \mu_{\mathbf{\Psi}} = \max\{\mu_{\mathbf{\Xi}}, \mu_{\mathbf{\Phi}}\} < 1 \ . \tag{18}$$

This expression of the multichannel mutual coherence is interesting as atoms can be selected based on their spatial or spectral morphology. In other words, discriminating two different multichannel atoms $\psi_{\gamma=\{i,p\}}$ and $\psi_{\gamma'=\{j,q\}}$ can be made based on:

- Spatial or temporal (resp. spectral) diversity : in this case $i = j$ and $p \neq q$ (resp. $i \neq j$ and $p = q$). These atoms have the same spectrum (resp. spatial shape) but one can discriminate between them based on their spatial (resp. spectral) diversity. From (18), their coherence is lower than $\mu_{\mathbf{\Phi}}$ (resp. $\mu_{\mathbf{\Xi}}$). Disentangling these multichannel atoms can equivalently be done in the monochannel case.

- Both diversities : $i \neq j$ and $p \neq q$, the "separation" task seems easier as the atoms don't share neither the same spectra nor the same spatial (or temporal) "shape". Note that from (18), the coherence between these atoms in this case is lower than $\mu_{\mathbf{\Xi}}\mu_{\mathbf{\Phi}} \leq \max\{\mu_{\mathbf{\Xi}}, \mu_{\mathbf{\Phi}}\}$.

Let us assume that the data $\mathbf{X}$ are K-sparse in $\mathbf{\Psi}$. Hence, $\mathbf{X}$ are the linear combination of K multichannel atoms:

$$\mathbf{X} = \sum_{\gamma \in \Lambda(X)} \alpha_\gamma \psi_\gamma \ , \tag{19}$$

[4] The adjectives *spectral* and *spatial* that characterize the dictionaries are not formal. Owing to the symmetry of the multichannel sparse decomposition problems, $\mathbf{\Xi}$ and $\mathbf{\Phi}$ have no formal difference. In practice and more particularly in multi/hyperspectral imaging, $\mathbf{\Xi}$ will refer to the dictionary of physical spectra and $\mathbf{\Phi}$ to the dictionary of image/signal waveforms.

This equation is clearly similar to the monochannel case. Owing to this key observation, we will see in the next paragraph that most sparse decomposition results can be extended to the multichannel case.

3.3.1 Multichannel sparse recovery results

In the last paragraph we emphasized on the apparent similarities between the monochannel and multichannel sparse models in Equation (19). Similarly, decomposing multichannel data in $\mathbf{\Psi}$ requires solving the following problem:

$$\min_{\alpha} \|\alpha\|_0 \text{ s.t } \mathbf{X} = \alpha\mathbf{\Psi} \ , \tag{20}$$

where $\alpha\mathbf{\Psi} = \sum_\gamma \alpha_\gamma \psi_\gamma$. The convex ℓ_1 minimization problem would be recast equivalently in the multichannel case:

$$\min_{\alpha} \|\alpha\|_1 \text{ s.t } \mathbf{X} = \alpha\mathbf{\Psi} \ . \tag{21}$$

From the optimization viewpoint, monochannel and multichannel problems are similar. This point leads us to straightforwardly extend sparse recovery results in Equation (16) to the multichannel case. The uniqueness and equivalence condition of the sparse multichannel decomposition problem in Equation (20) is then similar to the monochannel case. Assume that $\mathbf{X}$ is K-sparse in the multichannel dictionary $\mathbf{\Psi} = \mathbf{\Xi} \otimes \mathbf{\Phi}$. The ℓ_0 sparse decomposition problem in Equation (20) has a unique solution and problems in Equation (20) and (21) are equivalent when :

$$K < \frac{1}{2}\left(1 + \frac{1}{\mu_{\mathbf{\Psi}}}\right) \text{ where } \mu_{\mathbf{\Psi}} = \max\{\mu_{\mathbf{\Xi}}, \mu_{\mathbf{\Phi}}\} \ .$$

In this framework, most results in the monochannel case (49; 53; 51; 52; 1; 48) can be straightforwardly extended to the multichannel case.

3.3.2 Practical Sparse Signal Decomposition

In the previous sections, we emphasized on conditions under which the ℓ_0-sparse decomposition problem in Equation (20) can be replaced with the convex ℓ_1-sparse decomposition problem in Equation (21). Most algorithms that have been proposed to solve sparse decomposition issues can be divided into three main categories:

- Linear programming : in the seminal paper (47), the authors proposed to solve the convex ℓ_1-sparse decomposition problem in Equation (21) with linear programming methods such as interior point methods. Unfortunately, linear programming-based methods are computationally demanding and thus not well suited to large-scale problems such as ours.

- Greedy algorithms : the most popular greedy algorithm must be the Matching Pursuit and its orthogonal version OMP (59). Conditions have been given under which MP and OMP are proved to solve the ℓ_1 and ℓ_0 sparse decomposition problems (60; 1; 61). Greedy algorithms have also been proposed by the statistics community for solving variable selection problems (LARS/LASSO see (62; 63)). Homotopy-continuation algorithms have also been introduced to solve the sparse decomposition problem (64; 65; 66). Interestingly, a recent work by Donoho (67) enlightens the links between greedy algorithms such as OMP, variable selection algorithms and homotopy. Such greedy algorithms however suffer from high computational cost.
- Iterative thresholding : recently, iterative thresholding algorithms have been proposed to mitigate the greediness of the aforementioned *stepwise* algorithms. Iterative thresholding has first been introduced for solving sparsity-based inverse problems (see (68; 69; 70)).

Most of these algorithms can be easily extended to handle multichannel data.

4 Morphological Component Analysis For Multichannel Data

4.1 Morphological Diversity and Morphological Component Analysis

An introduction to morphological diversity

Recall that a monochannel signal x is said to be sparse in a waveform dictionary $\mathbf{\Phi}$ if it can be well represented from a few dictionary elements. As discussed in (37), a single basis is often not well-adapted to large classes of highly structured data such as "natural images". Furthermore, over the past ten years, new tools have emerged from modern computational harmonic analysis : wavelets (71), ridgelets (72), curvelets (73; 74; 44), bandlets (75), contourlets (76), to name a few. It is quite tempting to combine several representations to build a larger dictionary of waveforms that will enable the sparse representation of larger classes of signals.

In (37) and (77), the authors proposed a practical algorithm coined Morphological Component Analysis (MCA) aiming at decomposing signals in overcomplete dictionaries made of a union of bases. In the MCA setting, x is the linear combination of D morphological components:

$$x = \sum_{i=1}^{D} \varphi_i = \sum_{i=1}^{D} \alpha_i \mathbf{\Phi}_i \ , \tag{22}$$

where $\{\mathbf{\Phi}_i\}_{i=1,\cdots,D}$ are orthonormal bases of $\mathbb{R}^t$. Morphological diversity then relies on the sparsity of those morphological components in specific bases.

In terms of ℓ_0 quasi-norm, this morphological diversity can be formulated as follows:

$$\forall\{i,j\} \in \{1,\cdots,D\}; \quad j \neq i \Rightarrow \|\varphi_i \mathbf{\Phi}_i^T\|_{\ell_0} < \|\varphi_i \mathbf{\Phi}_j^T\|_{\ell_0} \ . \tag{23}$$

In other words, MCA relies on the incoherence between the sub-dictionaries $\{\mathbf{\Phi}_i\}_{i=1,\cdots,D}$ to estimate the morphological components $\{\varphi_i\}_{i=1,\cdots,D}$ by solving the following convex minimization problem:

$$\{\varphi_i\}_{1\leq i\leq D} = \underset{\{\varphi_i\}_{1\leq i\leq D}}{\arg\min} \left\| x - \sum_{i=1}^{D} \varphi_i \right\|_{\ell_2}^2 + 2\lambda \sum_{i=1}^{D} \|\varphi_i \mathbf{\Phi}_i^T\|_{\ell_1} \ . \tag{24}$$

Note that the minimization problem in (24) is closely related to Basis Pursuit Denoising (BPDN - see (47)). In (78), we proposed a particular block-coordinate relaxation, iterative thresholding algorithm (MCA/MOM) to solve (24). Theoretical arguments as well as experiments were given showing that MCA provides at least as good results as Basis Pursuit for sparse overcomplete decompositions in a union of bases. Moreover, MCA turns out to be clearly much faster than Basis Pursuit. Then, MCA is a practical alternative to classical sparse overcomplete decomposition techniques.

Morphological diversity in multichannel data

In the previous paragraph, we gave a brief description of morphological diversity in the monochannel case. We extend morphological diversity to the multichannel case. In this particular setting, we assume that each observation or channel $\{x_i\}_{i=1,\cdots,m}$ is the linear combination of D morphological components:

$$\forall i \in \{1,\cdots,m\}; \quad x_i = \sum_{j=1}^{D} \varphi_{ij} \ , \tag{25}$$

where each morphological component φ_{ij} is sparse in a specific basis $\mathbf{\Phi}_j$. Then each channel $\{x_i\}_{i=1,\cdots,m}$ is assumed to be sparse in the overcomplete dictionary $\mathbf{\Phi}$ made of the union of the D bases $\{\mathbf{\Phi}_i\}_{i=1,\cdots,D}$.
We further assume that each column of the data matrix $\mathbf{X}$ is sparse in the dictionary $\mathbf{\Xi}$ made of the union of D' bases $\{\mathbf{\Xi}_i\}_{i=1,\cdots,D'}$ to account for inter-channel structures. The multichannel data $\mathbf{X}$ are then assumed to be sparse in the multichannel dictionary $\mathbf{\Psi} = [\mathbf{\Xi_1} \cdots \mathbf{\Xi_{D'}}] \otimes [\mathbf{\Phi_1} \cdots \mathbf{\Phi_D}]$. The multichannel data are then modeled as the linear combination of $D \times D'$ multichannel morphological components:

$$\mathbf{X} = \sum_{j=1}^{D} \sum_{k=1}^{D'} \varpi_{jk} \ , \tag{26}$$

where ϖ_{jk} is sparse in $\mathbf{\Xi}_k \otimes \mathbf{\Phi}_j$. In the same vein as what we discussed in subsection 3.3 on how to discriminate two multichannel atoms, separating two multichannel morphological components ϖ_{ip} and $\varpi_{jq\neq ip}$ may be achieved based either on spatial/temporal (resp. spectral) morphologies ($i \neq j$ and $p = q$, resp. $i = j$ and $p \neq q$), or on both morphologies ($i \neq j$ and $p \neq q$). The "separation" task seems easier in the latter case as the morphological components share neither the same spectral basis nor the same spatial (or temporal) basis.

Analyzing multichannel signals requires accounting for their *spectral* and *spatial* morphological diversities. For that purpose, the proposed multichannel extension to MCA coined mMCA aims at solving the following minimization problem :

$$\min_{\{\varpi_{jk}\}} \left\| \mathbf{X} - \sum_{j=1}^{D} \sum_{k=1}^{D'} \varpi_{jk} \right\|_F^2 + 2\lambda \sum_{j=1}^{D} \sum_{k=1}^{D'} \|\mathbf{\Xi}_k^T \varpi_{jk} \mathbf{\Phi}_j^T\|_1 \ . \tag{27}$$

4.2 Multichannel overcomplete sparse recovery

General multichannel overcomplete sparse decomposition

Recall that $\mathbf{\Xi}$ is a $m \times M$ overcomplete dictionary with $M > m$, $\mathbf{\Phi}$ is a $T \times t$ overcomplete dictionary with $T > t$. Let us first consider the noiseless case. The multichannel extension of (13) writes as follows:

$$\min_{\alpha} \|\alpha\|_0 \text{ s.t } \mathbf{X} = \mathbf{\Xi}\alpha\mathbf{\Phi} \ , \tag{28}$$

where α is an $M \times T$ matrix (see also (20)). Arguing as in the monochannel case, the convex ℓ_1 minimization problem (14) can also be rewritten in the multichannel setting :

$$\min_{\alpha} \|\alpha\|_1 \text{ s.t } \mathbf{X} = \mathbf{\Xi}\alpha\mathbf{\Phi} \ , \tag{29}$$

see also (21).

4.3 Multichannel Morphological Component Analysis

The problem at stake in Equation (27) can be solved by extending to the multichannel case well-known sparse decomposition algorithms as reviewed in subsection 3.3.2. Extension of MP and OMP to the multichannel case has been proposed in (58). The aforementioned greedy methods iteratively select one dictionary atom at a time. Unfortunately, this *stepwise* selection of active atoms is burdensome and the process may be sped up as in (79) where

a faster *stagewise* Orthogonal Matching Pursuit (StOMP) is introduced. It is shown to solve the ℓ_0 sparse recovery problem in Equation (13) with random dictionaries under mild conditions.

Owing to the particular structure of the problem in Equation (27), extending the MCA algorithm (37) to the multichannel case would lead to faster and still effective decomposition results. Recall that in the mMCA setting, the data $\mathbf{X}$ are assumed to be the linear combination of $D \times D'$ morphological components $\{\varpi_{jk}\}_{j=1,\cdots,D;k=1,\cdots,D'}$. $\Lambda(\varpi_{jk})$ is the support of ϖ_{jk} in the subdictionary $\mathbf{\Psi}_{jk} = \mathbf{\Xi}_k \otimes \mathbf{\Phi}_j$. As $\mathbf{X}$ is K-sparse in the whole dictionary, $\sum_{j,k}\ \mathrm{Card}\left(\Lambda(\varpi_{jk})\right) = K$. The data can be decomposed as follows:

$$\mathbf{X} = \sum_{j=1}^{D}\sum_{k=1}^{D'} \varpi_{jk} = \sum_{j=1}^{D}\sum_{k=1}^{D'}\sum_{i\in\Lambda(\varpi_{jk})} \alpha_{jk}[i]\psi_{jk}[i]\ . \tag{30}$$

Substituting Equation (30) in Equation (27), the mMCA algorithm approaches the solution to Equation (27) by iteratively and alternately estimating each morphological component ϖ_{jk} in a Block-coordinate relaxed way (see (80)). Each matrix of coefficients α_{jk} is then updated as follows :

$$\alpha_{jk} = \arg\min_{\alpha_{jk}} \|\mathbf{R}_{jk} - \mathbf{\Xi}_k\ \alpha_{jk}\mathbf{\Phi}_j\|_F^2 + 2\lambda\|\alpha_{jk}\|_1\ , \tag{31}$$

where $\mathbf{R}_{jk} = \mathbf{X} - \sum_{p,q\neq j,k}\mathbf{\Xi}_q\alpha_{pq}\mathbf{\Phi}_p$ is a residual term.
Since we are assuming that the subdictionaries $\{\mathbf{\Phi}_j\}_j$ and $\{\mathbf{\Xi}_k\}_k$ are orthonormal, the update rule in Equation (31) is equivalent to the following:

$$\alpha_{jk} = \arg\min_{\alpha_{jk}} \left\|\mathbf{\Xi}_k^T\mathbf{R}_k\mathbf{\Phi}_j^T - \alpha_{jk}\right\|_F^2 + 2\lambda\|\alpha_{jk}\|_1\ , \tag{32}$$

which has a unique solution $\alpha_{jk} = \Delta_\lambda\left(\mathbf{\Xi}_k^T\mathbf{R}_k\mathbf{\Phi}_j^T\right)$ known as soft-thresholding with threshold λ as follows:

$$\Delta_\lambda(u[i]) = \begin{cases} 0 & \text{if } u[i] < \lambda \\ u[i] - \lambda\ \mathrm{sign}\ (u[i]) & \text{if } u[i] \geq \lambda \end{cases}\ . \tag{33}$$

For a fixed λ, mMCA selects groups of atoms based on their scalar product with the residual $\mathbf{R}_{jk}$. Assuming that we select only the most coherent atom (with the highest scalar product) with the residual $\mathbf{R}_{jk}$ then one mMCA iteration boils down to a stepwise multichannel Matching Pursuit (mMP) step. In contrast with mMP, the mMCA algorithm is allowed to select several atoms at each iteration. Thus, when hard-thresholding is used instead of soft-thresholding, mMCA is equivalent to a *stagewise* mMP algorithm. Allowing mMCA to select new atoms is obtained by decreasing the threshold λ at each iteration. The mMCA algorithm is summarized below:

1. Set the number of iterations $I_{\max}$ and threshold $\lambda^{(0)}$.
2. While $\lambda^{(h)}$ is higher than a given lower bound $\lambda_{\min}$ (*e.g.* can depend on the noise variance, see Section 4.5),

For $j = 1, \cdots, D$ and $k = 1, \cdots, D'$

- Compute the residual term $\mathbf{R}_{jk}^{(h)}$ assuming the current estimates of $\varpi_{pq \neq jk}$, $\tilde{\varpi}_{pq \neq jk}^{(h-1)}$ are fixed:
  $\mathbf{R}_{jk}^{(h)} = \mathbf{X} - \sum_{pq \neq jk} \tilde{\varpi}_{pq \neq jk}^{(h-1)}$.
- Estimate the current coefficients of $\tilde{\varpi}_{jk}^{(h)}$ by thresholding with threshold $\lambda^{(h)}$:
  $\tilde{\alpha}_{jk}^{(h)} = \Delta_{\lambda^{(h)}} \left(\mathbf{\Xi}_k^T \mathbf{R}_{jk}^{(h)} \mathbf{\Phi}_j^T \right)$.
- Get the new estimate of ϖ_{jk} by reconstructing from the selected coefficients $\tilde{\alpha}_{jk}^{(h)}$:
  $\tilde{\varpi}_{jk}^{(h)} = \mathbf{\Xi}_k \tilde{\alpha}_k^{(h)} \mathbf{\Phi}_j$.

3. Decrease the threshold $\lambda^{(h)}$ following a given strategy.

4.3.1 The thresholding strategy

In (78) we proposed a thresholding strategy that is likely to provide the solution to the ℓ_0 sparse monochannel problem. The strategy which goes by the name of MOM (for "Mean of Max") can be extended to the multichannel case. At each iteration h the residual is projected onto each sub-dictionary and we define :

$$m_{jk}^{(h-1)} = \left\| \mathbf{\Xi}_k^T \left(\mathbf{X} - \sum_{p,q} \mathbf{\Xi}_q \tilde{\alpha}_{pq}^{(h-1)} \mathbf{\Phi}_p \right) \mathbf{\Phi}_j^T \right\|_\infty . \tag{34}$$

The multichannel-MOM (mMOM) threshold is then computed as the mean of the two largest values in the set $\{m_{jk}^{(h-1)}\}_{j=1,\cdots,D;k=1,\cdots,D'}$

$$\lambda^{(h)} = \frac{1}{2} \left\{ m_{j_0 k_0}^{(h-1)} + m_{j_1 k_1}^{(h-1)} \right\} . \tag{35}$$

In the next section, we show conditions under which mMCA/mMOM selects atoms without error and converges asymptotically to the solution of the multichannel ℓ_0 sparse recovery problem in Equation (20).

4.4 Recovering sparse multichannel decompositions using mMCA

The mMOM rule defined in Equation (34)-(35) is such that mMCA will select, at each iteration, atoms belonging to the same subdictionary $\mathbf{\Psi}_{jk} = \mathbf{\Xi}_k \otimes \mathbf{\Phi}_j$. Although it seems more computationally demanding, the mMOM strategy has several nice properties. We show sufficient conditions under which i) mMCA/mMOM selects atoms belonging to the active atom set of the

solution of the ℓ_0 sparse recovery problem (Exact Selection Property), ii) mMCA/mMOM converges exponentially to $\mathbf{X}$ and its sparsest representation in $\mathbf{\Psi}$. Let us mention that the mMCA/mMOM exhibits an auto-stopping behavior, and requires only one parameter $\lambda_{\min}$ whose choice is easy and discussed in Section 4.5.
The next proposition states that mMCA/mMOM verifies the Exact Selection Property at each iteration.

Proposition 1 (Exact Selection Property) *Suppose that* $\mathbf{X}$ *is* K*-sparse such that :*

$$\mathbf{X} = \sum_{j=1}^{D} \sum_{k=1}^{D'} \sum_{i \in \Lambda(\varpi_{jk})} \alpha_{jk}[i] \psi_{jk}[i] \ ,$$

where $K = \sum_{j,k} Card\,(\Lambda(\varpi_{jk}))$ *satisfying* $K < \frac{\mu_{\Psi}^{-1}}{2}$*. At the* h*-th iteration, assume that the residual* $\mathbf{R}^{(h)}$ *is* K*-sparse such that :*

$$\mathbf{R}^{(h)} = \sum_{j=1}^{D} \sum_{k=1}^{D'} \sum_{i \in \Lambda(\varpi_{jk})} \beta_{jk}[i] \psi_{jk}[i] \ .$$

Then mMCA/mMOM picks up coefficients belonging to the support of $\mathbf{X}$ *at iteration* (h)*.*

When the previous Exact Selection Property holds, the next proposition shows that mMCA/mMOM converges exponentially to $\mathbf{X}$ and its sparsest representation in $\mathbf{\Psi} = [\mathbf{\Xi}_1 \cdots \mathbf{\Xi}_{D'}] \otimes [\mathbf{\Phi}_1 \cdots \mathbf{\Phi}_D]$.

Proposition 2 (Convergence) *Suppose that* $\mathbf{X}$ *is* K*-sparse such that :*

$$\mathbf{X} = \sum_{j=1}^{D} \sum_{k=1}^{D'} \sum_{i \in \Lambda(\varpi_{jk})} \alpha_{jk}[i] \psi_{jk}[i] \ ,$$

where $K = \sum_{j,k} Card\,(\Lambda(\varpi_{jk}))$*.*
If $K < \frac{\mu_{\Psi}^{-1}}{2}$ *then mMCA/mMOM converges exponentially to* $\mathbf{X}$ *and its sparsest representation in* $\mathbf{\Psi}$*. More precisely, the residual converges to zero at an exponential rate.*

See (81) for detailed proofs. Note that the above conditions are far from being sharp. Exact Selection and convergence may still be valid beyond the bounds retained in the latter two statements.

When bounded noise perturbs the data, the data are modeled as follows :

$$\mathbf{X} = \sum_{j=1}^{D} \sum_{k=1}^{D'} \sum_{i \in \Lambda(\varpi_{jk})} \alpha_{jk}[i] \psi_{jk}[i] + \mathbf{N} \tag{36}$$

where $\mathbf{N}$ is a bounded noise : $\|\mathbf{N}\|_F < \epsilon$. Sparse recovery then needs to solve the following problem:

$$\min_{\alpha_{jk}} \sum_{j=1}^{D} \sum_{k=1}^{D'} \|\alpha_{jk}\|_0 \text{ s.t } \left\| \mathbf{X} - \sum_{j=1}^{D} \sum_{k=1}^{D'} \mathbf{\Xi}_k \alpha_{jk} \mathbf{\Phi}_j \right\|_F < \epsilon \tag{37}$$

Stability conditions of sparse recovery have been investigated in (82; 83; 84) in the monochannel case. More particularly, conditions are proved in (82) under which OMP verifies an Exact Selection Property in the presence of bounded noise. They also showed that the OMP solution lies in a ℓ_2 ball centered on the exact solution to the ℓ_0 sparse recovery problem with a radius on the order of ϵ. Exhibiting similar stability results in the mMCA setting is challenging and will be addressed in the future. In the mMCA framework, assuming the noise level is known, the mMCA/mMOM algorithm stops when $\lambda \leq \lambda_{\min}$ with $\lambda_{\min} = 3 - 4\epsilon$.

4.6 Choosing the overcomplete dictionary

The choice of the overcomplete dictionary is a key step as it determines where we will be looking for a sparse representation. It is the expression of some prior information we have available on the signal. Interestingly, the ℓ_1 sparse recovery problem can be seen in the light of a Bayesian framework. Solving the following problem

$$\min_{\{\alpha_{jk}\}} \left\| \mathbf{X} - \sum_{j=1}^{D} \sum_{k=1}^{D'} \mathbf{\Xi}_k \alpha_{jk} \mathbf{\Phi}_j \right\|_F^2 + 2\lambda \sum_{j=1}^{D} \sum_{k=1}^{D'} \|\alpha_{jk}\|_1 \tag{38}$$

is equivalent, in a Bayesian framework, to making the assumption among others of an independent Laplacian *prior* on the coefficients of each morphological component in the sparse representation domain. Choosing the set of subdictionaries is then equivalent to assuming some specific *prior* for each morphological component.
Furthermore, the attractiveness of mMCA lies in its ability to take advantage

of sparse representations which have fast implicit analysis and synthesis operators without requiring the explicit manipulation of each atom: wavelets (71), curvelets (74), bandlets (75), contourlets (76), ridgelets (72), wave atoms (85) to name a few. As a consequence, mMCA is a fast non-linear sparse decomposition algorithm whose computational complexity is dominated by that of the transforms involved in the dictionary.
In the image processing experiments reported in this paper, we will assume that a wide range of images can be decomposed into a piecewise smooth (contour) part and an oscillating texture part. We will assume *a priori* that the contour part is sparse in the curvelet tight frame, and the texture part is sparsely described by the local discrete cosine transform (DCT) (71) [5] . However, all the results we previously proved were given assuming that each subdictionary was an orthonormal basis. When the selected subdictionaries are more generally tight frames, the solution to (32) is no longer a simple thresholding. Nevertheless, in (86) and (70), the authors showed that thresholding is the first step towards solving (32) when the subdictionary is redundant. Rigorously, proximal-type iterative shrinkage is shown to converge to a solution of (32). In practice, even when the subdictionary is a tight frame (for instance the curvelet frame) we will only use a single thresholding step to solve (32).
As far as the choice of the *spectral* dictionary $\mathbf{\Xi}$ is concerned, it is based on a spectral sparsity assumption.

Epilogue

In this Section, we have surveyed the tricky problem raised by sparse overcomplete signal decomposition for multichannel data. We then presented a multichannel extension to the MCA algorithm. The so-called mMCA algorithm will be the backbone of the next sparsity-based algorithm we propose to solve the sparse BSS issue.

5 Morphological Diversity and Blind Source Separation

In (35) we introduced an extension of the mMCA framework for BSS. The so-called Generalized Morphological Component Analysis (GMCA) framework states that the observed data $\mathbf{X}$ are generated according to Equation (2). In words, $\mathbf{X}$ is a linear instantaneous mixture of unknown sources $\mathbf{S}$ using an unknown mixing matrix $\mathbf{A}$, with an additive perturbation term $\mathbf{N}$ that accounts for noise or model imperfection. We remind the reader that we only

[5] An alternative choice would be the wave atoms (85).

consider the overdetermined source separation case, *i.e.* $m \geq n$ and thus $\mathbf{A}$ has full column rank.

5.1 Generalized Morphological Component Analysis

From now, we assume that the sources are sparse in the *spatial* dictionary $\mathbf{\Phi}$ that is the concatenation of D orthonormal bases $\{\mathbf{\Phi}_i\}_{i=1,\cdots,D}$: $\mathbf{\Phi} = \left[\mathbf{\Phi}_1^T, \cdots, \mathbf{\Phi}_D^T\right]^T$. In the GMCA setting, each source is modeled as the linear combination of D morphological components where each component is sparse in a specific basis :

$$\forall i \in \{1, \cdots, n\}; \qquad s_i = \sum_{k=1}^{D} \varphi_{ik} = \sum_{k=1}^{D} \alpha_{ik} \mathbf{\Phi}_k \ . \tag{39}$$

GMCA seeks an unmixing scheme, through the estimation of $\mathbf{A}$, which leads to the sparsest sources $\mathbf{S}$ in the dictionary $\mathbf{\Phi}$. This is expressed by the following optimization task written in its augmented Lagrangian form:

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} 2\lambda \sum_{i=1}^{n} \sum_{k=1}^{D} \|\varphi_{ik} \mathbf{\Phi}_k^T\|_0 + \|\mathbf{X} - \mathbf{AS}\|_F^2 \ , \tag{40}$$

where each row of $\mathbf{S}$ is such that $s_i = \sum_{k=1}^{D} \varphi_{ik}$. Obviously this algorithm is combinatorial by nature. We then propose to substitute the ℓ_1 norm for the ℓ_0 sparsity, which amounts to solving the optimization problem :

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} 2\lambda \sum_{i=1}^{n} \sum_{k=1}^{D} \|\varphi_{ik} \mathbf{\Phi}_k^T\|_1 + \|\mathbf{X} - \mathbf{AS}\|_F^2 \ . \tag{41}$$

More conveniently, the product $\mathbf{AS}$ can be split into $n \times D$ multichannel morphological components: $\mathbf{AS} = \sum_{i,k} a^i \varphi_{ik}$. Based on this decomposition, we propose an alternating minimization algorithm to estimate iteratively one term at a time. Define the $\{i,k\}$-th multichannel residual by $\mathbf{R}_{i,k} = \mathbf{X} - \sum_{\{p,q\} \neq \{i,k\}} a^p \varphi_{pq}$ as the part of the data $\mathbf{X}$ unexplained by the multichannel morphological component $a^i \varphi_{ik}$. Estimating the morphological component $\varphi_{ik} = \alpha_{ik} \mathbf{\Phi}_k$ assuming $\mathbf{A}$ and $\varphi_{\{pq\} \neq \{ik\}}$ are fixed leads to the component-wise optimization problem :

$$\tilde{\varphi}_{ik} = \arg\min_{\varphi_{ik}} 2\lambda \|\varphi_{ik} \mathbf{\Phi}_k^T\|_1 + \|\mathbf{R}_{i,k} - a^i \varphi_{ik}\|_F^2 \ , \tag{42}$$

or equivalently,

$$\tilde{\alpha}_{ik} = \arg\min_{\alpha_{ik}} 2\lambda \|\alpha_{ik}\|_1 + \|\mathbf{R}_{i,k} \mathbf{\Phi}_k^T - a^i \alpha_{ik}\|_F^2 \ , \tag{43}$$

since here $\mathbf{\Phi}_k$ is an orthogonal matrix. By classical ideas in convex analysis, a necessary condition for $\tilde{\alpha}_{ik}$ to be a minimizer of the above functional is that

the null vector be an element of its subdifferential at $\tilde{\alpha}_{ik}$, that is :

$$0 \in -\frac{1}{\|a^i\|_2^2} a^{i^T} \mathbf{R}_{i,k} \mathbf{\Phi}_k^T + \alpha_{ik} + \frac{\lambda}{\|a^i\|_2^2} \partial \|\alpha_{ik}\|_1 \ , \tag{44}$$

where $\partial \|\alpha_{ik}\|_1$ is the subgradient defined as (owing to the separability of the ℓ_1-norm):

$$\partial \|\alpha\|_1 = \left\{ u \in \mathbb{R}^t \middle| \begin{array}{ll} u[l] = & \text{sign}(\alpha[l]), \quad l \in \Lambda(\alpha) \\ u[l] & \in [-1,1], \quad \text{otherwise.} \end{array} \right\}.$$

Hence, (44) can be rewritten equivalently as two conditions leading to the following closed-form solution:

$$\hat{\alpha}_{jk}[l] = \begin{cases} 0, & \text{if } \left| \left(a^{i^T} \mathbf{X}_{i,k} \mathbf{\Phi}_k^T \right)[l] \right| \leq \lambda \\ \alpha'[l], & \text{otherwise.} \end{cases} \tag{45}$$

where $\alpha' = \frac{1}{\|a^i\|_2^2} a^{i^T} \mathbf{R}_{i,k} \mathbf{\Phi}_k^T - \frac{\lambda}{\|a^i\|_2^2} \text{sign}\left(a^{i^T} \mathbf{R}_{i,k} \mathbf{\Phi}_k^T \right)$. This exact solution is known as soft-thresholding. Hence, the closed-form estimate of the morphological component φ_{ik} is:

$$\tilde{\varphi}_{ik} = \Delta_\delta \left(\frac{1}{\|a^i\|_2^2} a^{i^T} \mathbf{X}_{i,k} \mathbf{\Phi}_k^T \right) \mathbf{\Phi}_k \text{ with } \delta = \frac{\lambda}{\|a^i\|_2^2} \ . \tag{46}$$

Now, considering fixed $\{a^p\}_{p \neq i}$ and $\mathbf{S}$, updating the column a^i is then just a least-squares estimate:

$$\tilde{a}^i = \frac{1}{\|s_i\|_2^2} \left(\mathbf{X} - \sum_{p \neq i} a^p s_p \right) s_i^T \tag{47}$$

where $s_k = \sum_{k=1}^D \varphi_{ik}$. In a simpler context, this iterative and alternating optimization scheme has already proved its efficiency in (34).

In practice each column of $\mathbf{A}$ is forced to have unit ℓ_2 norm at each iteration to avoid the classical scale indeterminacy of the product $\mathbf{AS}$ in Equation (2). The GMCA algorithm is summarized below:

1. Set the number of iterations $I_{\max}$ and threshold $\delta^{(0)}$
2. While $\delta^{(h)}$ is higher than a given lower bound $\delta_{\min}$ (*e.g.* can depend on the noise standard deviation),
For $i = 1, \cdots, n$
For $k = 1, \cdots, D$
• Compute the term $r_{ik}^{(h)}$ assuming the current estimates of $\varphi_{\{pq\}\neq\{ik\}}$, $\tilde{\varphi}_{\{pq\}\neq\{ik\}}^{(h-1)}$ are fixed :
$$r_{ik}^{(h)} = \tilde{a}^{i^{(h-1)^T}} \left(\mathbf{X} - \sum_{\{p,q\}\neq\{i,k\}} \tilde{a}^{p^{(h-1)}} \tilde{\varphi}_{\{pq\}}^{(h-1)} \right)$$
• Estimate the current coefficients of $\tilde{\varphi}_{ik}^{(h)}$ by thresholding with threshold $\delta^{(h)}$:
$$\tilde{\alpha}_{ik}^{(h)} = \Delta_{\delta^{(h)}} \left(r_{ik}^{(h)} \mathbf{\Phi}_k^T \right)$$
• Get the new estimate of φ_{ik} by reconstructing from the selected coefficients $\tilde{\alpha}_{ik}^{(h)}$:
$$\tilde{\varphi}_{ik}^{(h)} = \tilde{\alpha}_{ik}^{(h)} \mathbf{\Phi}_k$$
Update a^i assuming $a^{p\neq k^{(h)}}$ and the morphological components $\tilde{\varphi}_{pq}^{(h)}$ are fixed:
$$\tilde{a}^{i^{(h)}} = \frac{1}{\|\tilde{s}_i^{(h)}\|_2^2} \left(\mathbf{X} - \sum_{p\neq i}^{n} \tilde{a}^{p^{(h-1)}} \tilde{s}_p^{(h)} \right) \tilde{s}_i^{(h)^T}$$
– Decrease the threshold $\delta^{(h)}$.

GMCA is an iterative thresholding algorithm where at each iteration, *coarse* versions of the morphological component $\{\varphi_{ik}\}_{i=1,\cdots,n;k=1,\cdots,D}$ for each source s_i are first computed. These raw sources are estimated from their most significant coefficients in $\mathbf{\Phi}$. This first step then amounts to performing a single mMCA decomposition step in the multichannel representation $\mathbf{A} \otimes \mathbf{\Phi}$ with the threshold $\delta^{(h)}$.
Following this step, the column a^i corresponding to the i-th source is estimated from the most significant features of s_i. Each source and its corresponding column of $\mathbf{A}$ are then alternately estimated. The whole optimization scheme then progressively refines the estimates of $\mathbf{S}$ and $\mathbf{A}$ as δ decreases towards $\delta_{\min}$. This particular iterative thresholding scheme provides robustness to the algorithm by working first on the most significant features in the data and then progressively incorporating smaller details to finely tune the model parameters. The main difference with the mMCA algorithm lies in the mixing matrix update. Such stage is then equivalent to updating a part of the multichannel dictionary in which mMCA decomposes the data $\mathbf{X}$.

5.1.1 The dictionary $\mathbf{\Phi}$

As an MCA-like algorithm (see (77)), the GMCA algorithm involves multiplications by matrices $\mathbf{\Phi}_k^T$ and $\mathbf{\Phi}_k$. Thus, GMCA is attractive in large-scale problems as long as the redundant dictionary $\mathbf{\Phi}$ is a union of bases or tight frames. For such dictionaries, matrices $\mathbf{\Phi}_k^T$ and $\mathbf{\Phi}_k$ are never explicitly constructed,

and fast implicit analysis and reconstruction operators are used instead (for instance, wavelet transforms, global or local discrete cosine transform, etc).

5.1.2 Complexity analysis

We here provide a detailed analysis of the complexity of GMCA. We begin by noting that the bulk of the computation is invested in the application of $\mathbf{\Phi}_k^T$ and $\mathbf{\Phi}_k$ at each iteration and for each component. Hence, fast implicit operators associated to $\mathbf{\Phi}_k$ or its adjoint are of key importance in large-scale applications. In our analysis below, we let V_k denote the cost of one application of a linear operator $\mathbf{\Phi}_k$ or its adjoint. The computation of the multichannel residuals for all (i, k) costs $O(nDmt)$ flops. Each step of the double 'For' loop computes the correlation of this residual with a^{i^T} using $O(mt)$ flops. Next, it computes the residual correlations (application of $\mathbf{\Phi}_k^T$), thresholds them, and then reconstructs the morphological component φ_{ik}. This costs $O(2V_k + T)$ flops. The sources are then reconstructed with $O(nDt)$, and the update of each mixing matrix column involves $O(mt)$ flops. Noting that in our setting, $n \sim m \ll t$, and $V_k = O(t)$ or $O(t \log t)$ for most popular transforms, the whole GMCA algorithms then costs $O(I_{\max} n^2 Dt) + O(2I_{\max} n \sum_{k=1}^{D} V_k + nDT)$. Thus, in practice GMCA could be computationally demanding for large scale high dimensional problems. In Section 5.3, we prove that adding some more assumptions leads to a very simple, accurate and much faster algorithm that enables to handle very large scale problems.

5.1.3 The thresholding strategy

Hard or Soft-thresholding ? Rigorously, we should use a soft-thresholding operator. In practice, hard-thresholding leads to better results. Furthermore in (78), it was shown empirically that the use of hard-thresholding is likely to provide the ℓ_0 sparse solution for the single channel sparse decomposition problem. By analogy, the use of a hard-thresholding operator is assumed to solve the multichannel ℓ_0 quasi-norm problem instead of (41).

Handling noise The GMCA algorithm is well suited to deal with noisy data. Assume that the noise standard deviation is $\sigma_{\mathbf{N}}$. Then, we simply apply the GMCA algorithm as described above, terminating as soon as the threshold δ gets less than $\tau\sigma_{\mathbf{N}}$. Here, τ typically takes its value in the range $3 - 4$. This attribute of GMCA makes it a suitable choice for use in noisy applications. GMCA not only manages to separate the sources, but also succeeds in removing additive noise as a by-product.

5.1.4 The Bayesian point of view

We can also consider GMCA from a Bayesian viewpoint. For instance, let's assume that the entries of the mixtures $\{x_i\}_{i=1,\cdots,m}$, the mixing matrix $\mathbf{A}$, the sources $\{s_j\}_{j=1,\cdots,n}$ and the noise matrix $\mathbf{N}$ are random variables. For simplicity, $\mathbf{N}$ is Gaussian; its samples are *iid* from a multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{\Sigma_N})$ with zero mean and covariance matrix $\mathbf{\Sigma_N}$. The noise covariance matrix $\mathbf{\Sigma_N}$ is assumed known. For simplicity, the noise samples are considered to be decorrelated from one channel to the other; the covariance matrix $\mathbf{\Sigma_N}$ is thus diagonal. We assume that each entry of $\mathbf{A}$ is generated from a uniform distribution. Let us remark that other priors on $\mathbf{A}$ could be imposed here; *e.g.* known fixed column for example.
We assume that the sources $\{s_i\}_{i=1,\cdots,n}$ are statistically independent from each other and their coefficients in $\mathbf{\Phi}$ (the $\{\alpha_i\}_{i=1,\cdots,n}$) are generated from a Laplacian law:

$$\forall i = 1, \cdots, n; \qquad f_{\mathbf{S}}(\alpha_i) = \prod_{k=1}^{T} f_{\mathbf{S}}(\alpha_i[k]) \propto \exp\left(-\mu \|\alpha_i\|_1\right) \ . \tag{48}$$

In a Bayesian framework, the use of the Maximum a posteriori estimator leads to the following optimization problem:

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} \|\mathbf{X} - \mathbf{AS}\|^2_{\mathbf{\Sigma_N}} + 2\mu \sum_{i=1}^{n} \sum_{k=1}^{D} \|\varphi_{ik} \mathbf{\Phi}_k^T\|_1 \ , \tag{49}$$

where $\|.\|_{\mathbf{\Sigma_N}}$ is the Frobenius norm defined by $\|\mathbf{X}\|^2_{\mathbf{\Sigma_N}} = \text{Trace}\left(\mathbf{X^T \Sigma_N^{-1} X}\right)$. Note that this minimization task is similar to (41) except that here the data fidelity term involving the norm $\|.\|_{\mathbf{\Sigma_N}}$ accounts for noise. In the case of homoscedastic and decorrelated noise (*i.e.* $\mathbf{\Sigma_N} = \sigma^2_{\mathbf{N}} \mathbf{I}_m$), problems (41) and (49) are equivalent with $\lambda = \mu \sigma^2_{\mathbf{N}}$. Note that in this framework the independence assumption in Equation (48) does not necessarily entail that the sources are "truly" independent. Rather it means that there are no *a priori* assumptions that indicate any dependency between the sources.

5.2 Results

We illustrate here the performance of GMCA with a simple toy experiment. We consider two sources s_1 and s_2 sparse in the union of the DCT and a discrete orthonormal wavelet basis. Their coefficients in $\mathbf{\Phi}$ are randomly generated from a Bernoulli-Gaussian distribution: the probability for a coefficient $\{\alpha_{1,2}[k]\}_{k=1,\cdots,T}$ to be non-zero is $p = 0.01$ and its amplitude is drawn from a Gaussian distribution with mean 0 and variance 1. The signals were composed of $t = 1024$ samples. We define the mixing matrix criterion

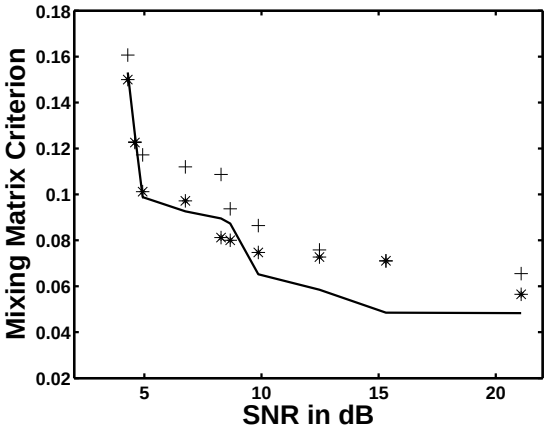

Fig. 2. Evolution of the mixing matrix criterion Δ_A as the noise variance varies: GMCA (*solid* line), EFICA, *(⋆) :*, RNA *(+)*. **Abscissa :** SNR in dB. **Ordinate :** mixing matrix criterion value.

$\mathcal{C}_{\mathbf{A}} = \|\mathbf{I}-\mathbf{P}\hat{\mathbf{A}}^{-1}\mathbf{A}\|_{1,1}$, where $\mathbf{P}$ is a matrix that reduces the scale/permutation indeterminacy of the mixing model. Indeed, when $\mathbf{A}$ is perfectly estimated, it is equal to $\hat{\mathbf{A}}$ up to scale and permutation. In the simulation experiments, the true sources and mixing matrix are obviously known and thus $\mathbf{P}$ can be computed easily. The mixing matrix criterion is thus strictly positive unless the mixing matrix is perfectly estimated up to scale and permutation. This mixing matrix criterion is experimentally much more sensitive to separation errors. Figure 2 illustrates the evolution of $\mathcal{C}_{\mathbf{A}}$ as the Signal-to-Noise Ratio $\mathrm{SNR} = 10\log_{10}\left(\|\mathbf{AS}\|_2^2/\|\mathbf{N}\|_2^2\right)$ increases. We compare our method to the Relative Newton Algorithm (RNA) (29) that accounts for sparsity and EFICA (21). The latter is a FastICA variant designed for highly leptokurtic sources. Both RNA and EFICA were applied after "sparsifying" the data via an orthonormal wavelet transform. Figure 2 shows that GMCA behaves similarly to state-of-the-art sparse BSS techniques.

5.3 Speeding up blind-GMCA

5.3.1 Introduction: the orthonormal case

Let us assume that the dictionary $\mathbf{\Phi}$ is no longer redundant and reduces to an orthonormal basis. The ℓ_0 optimization problem (40) then boils down to the following one:

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} \|\mathbf{\Theta}_{\mathbf{X}} - \mathbf{A}\alpha\|_F^2 + 2\lambda \sum_{i=1}^{n} \|\alpha_i\|_0 \quad \text{with } \mathbf{S} = \alpha\mathbf{\Phi}, \qquad (50)$$

where each row of $\mathbf{\Theta_X} = \mathbf{X}\mathbf{\Phi}^T$ stores the decomposition of each observed channel in $\mathbf{\Phi}$. Similarly the ℓ_1 norm problem (41) reduces to :

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} \|\mathbf{\Theta_X} - \mathbf{A}\alpha\|_F^2 + 2\lambda \sum_{i=1}^{n} \|\alpha_i\|_1 \quad \text{with } \mathbf{S} = \alpha\mathbf{\Phi} \ . \tag{51}$$

The GMCA algorithm no longer needs transforms at each iteration as only the data $\mathbf{X}$ have to be transformed once in $\mathbf{\Phi}$. Clearly, this case is computationally much cheaper. Unfortunately, no orthonormal basis is able to sparsely represent large classes of signals and yet we would like to use "very" sparse signal representations which motivated the use of redundant representations in the first place. The next section gives a few arguments supporting the substitution of (51) for (41) even when the dictionary $\mathbf{\Phi}$ is redundant.

The redundant case In this section, we assume $\mathbf{\Phi}$ is redundant. We consider that each datum $\{x_i\}_{i=1,\cdots,m}$ has a unique ℓ_0 sparse decomposition (*i.e.* $\mathcal{S}_{\ell_0}^{\mathbf{\Phi}}(x_i)$ is a singleton for any $i \in \{1, \cdots, m\}$). We also assume that the sources have unique ℓ_0 sparse decompositions (*i.e.* $\mathcal{S}_{\ell_0}^{\mathbf{\Phi}}(s_i)$ is a singleton for all $i \in \{1, \cdots, n\}$). We then define $\mathbf{\Theta_X} = [\Delta_{\mathbf{\Phi}}(x_1)^T, \cdots, \Delta_{\mathbf{\Phi}}(x_m)^T]^T$ and $\mathbf{\Theta_S} = [\Delta_{\mathbf{\Phi}}(s_1)^T, \cdots, \Delta_{\mathbf{\Phi}}(s_n)^T]^T$.
Up to now, we believed in morphological diversity as the source of discernibility between the sources we wish to separate. Thus, distinguishable sources must have " discernibly different" supports in $\mathbf{\Phi}$. Intuition then tells us that *when one mixes very sparse sources their mixtures should be less sparse*. Two cases have to be considered:

- Sources with disjoint supports in $\mathbf{\Phi}$: the mixing process increases the ℓ_0 norm : $\|\Delta_{\mathbf{\Phi}}(x_j)\|_{\ell_0} > \|\Delta_{\mathbf{\Phi}}(s_i)\|_{\ell_0}$ for all $j \in \{1, \cdots, m\}$ and $i \in \{1, \cdots, n\}$. When $\mathbf{\Phi}$ is made of a single orthogonal basis, this property is exact.
- Sources with δ-disjoint supports in $\mathbf{\Phi}$: the argument is not so obvious; we conjecture that the number of significant coefficients in $\mathbf{\Phi}$ is higher for mixture signals than for the original sparse sources with high probability : $\text{Card}\,(\Lambda_\delta(x_j)) > \text{Card}\,(\Lambda_\delta(s_i))$ for any $j \in \{1, \cdots, m\}$ and $i \in \{1, \cdots, n\}$.

Owing to this "intuitive" viewpoint, even in the redundant case, the method is likely to solve the following optimization problem :

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{\Theta}}_{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{\Theta_S}} \|\mathbf{\Theta_X} - \mathbf{A}\mathbf{\Theta_S}\|_F^2 + 2\lambda\|\mathbf{\Theta_S}\|_0 \ . \tag{52}$$

Obviously, (52) and (40) are not equivalent unless $\mathbf{\Phi}$ is orthonormal. When $\mathbf{\Phi}$ is redundant, no rigorous mathematical proof is easy to derive. Nevertheless, experiments will show that intuition leads to good results. In (52), note that a key point is still doubtful : sparse redundant decompositions (operator $\Delta_{\mathbf{\Phi}}$) are non-linear and in general no linear model is preserved. Writing $\Delta_{\mathbf{\Phi}}(\mathbf{\Theta_X}) =$

$\mathbf{A}\Delta_{\mathbf{\Phi}}(\mathbf{\Theta_S})$ at the solution is then an invalid statement in general. The next section will focus on this source of fallacy.

When non-linear processes preserve linearity Whatever the sparse decomposition used (*e.g.* Matching Pursuit (87), Basis Pursuit (47), etc), the decomposition process is non-linear. The simplification we made earlier is no longer valid unless the decomposition process preserves linear mixtures. Let us first focus on a single signal : assume that y is the linear combination of m original signals (y could be a single datum in the BSS model) :

$$y = \sum_{i=1}^{m} \nu_i y_i \ . \tag{53}$$

Assuming each $\{y_i\}_{i=1,\cdots,m}$ has a unique ℓ_0 sparse decomposition, we define $\alpha_i = \Delta_{\mathbf{\Phi}}(y_i)$ for all $i \in \{1, \cdots, m\}$. As defined earlier, $\mathcal{S}_{\ell_0}^{\mathbf{\Phi}}(y)$ is the set of ℓ_0 sparse solutions perfectly synthesizing y: for any $\alpha \in \mathcal{S}_{\ell_0}^{\mathbf{\Phi}}(y)$; $\quad y = \alpha\mathbf{\Phi}$. Amongst these solutions, one is the linearity-preserving solution $\alpha^\star$ defined such that:

$$\alpha^\star = \sum_{i=1}^{m} \nu_i \alpha_i \ . \tag{54}$$

As $\alpha^\star$ belongs to $\mathcal{S}_{\ell_0}^{\mathbf{\Phi}}(y)$, a sufficient condition for the ℓ_0 sparse decomposition to preserve linearity is the uniqueness of the sparse decomposition. Indeed, (46) proved that, in the general case, if

$$||\alpha||_0 < (\mu_{\mathbf{\Phi}}^{-1} + 1)/2 \ , \tag{55}$$

then this is the unique maximally sparse decomposition, and that in this case $\mathcal{S}_{\ell_1}^{\mathbf{\Phi}}(y)$ contains this unique solution as well. Therefore, if all the sources have sparse enough decompositions in $\mathbf{\Phi}$ in the sense of inequality (55), then the sparse decomposition operator $\Delta_{\mathbf{\Phi}}(.)$ preserves linearity.

In (78), the authors showed that when $\mathbf{\Phi}$ is the union of D orthonormal bases, MCA is likely to provide the unique ℓ_0 pseudo-norm sparse solution to the problem (13) under the assumption that the sources are sparse enough. Furthermore, in (78), experiments illustrate that the uniqueness bound (55) is too pessimistic. Uniqueness should hold, with high probability, beyond the bound (55). Hence, based on this discussion and the results reported in (78), we consider in the next experiments that the operation $\Delta_{\mathbf{\Phi}}(y)$ which stands for the decomposition of y in $\mathbf{\Phi}$ using MCA, preserves linearity.

In the BSS context In the BSS framework, recall that each observation $\{x_i\}_{i=1,\cdots,m}$ is the linear combination of n sources :

$$x_i = \sum_{j=1}^{n} a_{ij} s_j \ . \tag{56}$$

Owing to the last paragraph, if the sources and the observations have unique ℓ_0-sparse decompositions in $\mathbf{\Phi}$ then the linear mixing model is preserved, that is:

$$\Delta_{\mathbf{\Phi}}(x_i) = \sum_{j=1}^{n} a_{ij} \Delta_{\mathbf{\Phi}}(s_j) \ , \tag{57}$$

and we can estimate both the mixing matrix and the sources in the sparse domain by solving (52).

5.3.2 The Fast blind GMCA algorithm

According to the last section, a fast GMCA algorithm working in the sparse transform domain (after decomposing the data in $\mathbf{\Phi}$ using a sparse decomposition algorithm) could be designed to solve (50) (respectively (51)) by an iterative and alternate estimation of $\mathbf{\Theta_S}$ and $\mathbf{A}$. There is an additional important simplification when substituting problem (51) for (41). Indeed, as $m \geq n$, it turns out that (51) is a multichannel *overdetermined* least-squares error fit with ℓ_1-sparsity penalization. We again use an alternating minimization scheme to solve for $\mathbf{A}$ and $\mathbf{\Theta_S}$:

- Update the coefficients: when $\mathbf{A}$ is fixed, the marginal optimization problem has a unique solution given by the forward-backward proximal fixed-point equation, see (70, Proposition 3.1):

$$\tilde{\mathbf{\Theta}}_{\mathbf{S}} = \Delta_\delta \left(\tilde{\mathbf{\Theta}}_{\mathbf{S}} + \mathbf{M}(\mathbf{\Theta_X} - \mathbf{A}\tilde{\mathbf{\Theta}}_{\mathbf{S}}) \right) \tag{58}$$

  where $\mathbf{M}$ is a relaxation descent-direction matrix such that the spectral radius of $\mathbf{I} - \mathbf{MA}$ is bounded above by 1. Choosing $\mathbf{M} = \tilde{\mathbf{A}}^\dagger$ (pseudo-inverse of $\mathbf{A}$ which is full column-rank), gives $\tilde{\mathbf{\Theta}}_{\mathbf{S}} = \Delta_\delta \left(\tilde{\mathbf{A}}^\dagger \mathbf{\Theta_X} \right)$, Δ_δ is a thresholding operator (hard for (50) and soft for (51)) and the threshold δ decreases with increasing iteration count assuming.
- Update the mixing matrix $\mathbf{A}$ by a least-squares estimate: $\tilde{\mathbf{A}} = \mathbf{\Theta_X} \tilde{\mathbf{\Theta}}_{\mathbf{S}}^{\mathbf{T}} \left(\tilde{\mathbf{\Theta}}_{\mathbf{S}} \tilde{\mathbf{\Theta}}_{\mathbf{S}}^{\mathbf{T}} \right)^{-1}$.

Note that the latter two step estimation scheme has the flavor of the alternating *Sparse coding/Dictionary learning* algorithm presented in (88) in a different framework.

The two stages iterative process leads to the following fast GMCA algorithm:

1. Perform a MCA to each data channel to compute $\mathbf{\Theta_X}$:
   $\mathbf{\Theta_X} = [\Delta_{\mathbf{\Phi}}(x_i)^T]^T$.
2. Set the number of iterations $I_{\max}$ and threshold $\delta^{(0)}$.
3. While each $\delta^{(h)}$ is higher than a given lower bound $\delta_{\min}$ (*e.g.* can depend on the noise standard deviation),
   – Proceed with the following iteration to estimate the coefficients of the sources $\mathbf{\Theta_S}$ at iteration h assuming $\mathbf{A}$ is fixed:
   $\mathbf{\Theta_S}^{(h+1)} = \Delta_{\delta^{(h)}}\left(\mathbf{A}^{\dagger(h)}\mathbf{\Theta_X}\right)$.
   – Update $\mathbf{A}$ assuming $\mathbf{\Theta_S}$ is fixed :
   $\tilde{\mathbf{A}}^{(h+1)} = \mathbf{\Theta_X}\tilde{\mathbf{\Theta}}_{\mathbf{S}}^{(h)^T}\left(\tilde{\mathbf{\Theta}}_{\mathbf{S}}^{(h)}\tilde{\mathbf{\Theta}}_{\mathbf{S}}^{(h)^T}\right)^{-1}$.
   – Decrease the threshold $\delta^{(h)}$.
4. Stop when $\delta^{(h)} = \delta_{\min}$.

In the same vein as in subsection 5.1, the *coarse to fine* process is also the core of this fast version of GMCA. Indeed, when $\delta^{(h)}$ is high, the sources are estimated from their most significant coefficients in $\mathbf{\Phi}$. Intuitively, the coefficients with high amplitude in $\mathbf{\Theta_S}$ are (i) less perturbed by noise and (ii) should belong to only one source with overwhelming probability. The estimation of the sources is refined as the threshold δ decreases towards a final value $\delta_{\min}$. Similarly to the previous version of the GMCA algorithm, the optimization process provides robustness to noise and helps convergence even in a noisy context. Experiments in Section 5.6 illustrate the good performances of this fast GMCA algorithm.

Complexity analysis When the approximations we make are valid, the fast simplified GMCA version requires only the application of MCA on each channel, which is faster than the non-fast version (see subsection 5.1.2). Indeed, once MCA is applied on each channel, the rest of the algorithm requires $O(I_{\max}n^2Dt)$ flops.

5.3.3 A fixed point algorithm

Recall that the GMCA algorithm is composed of two steps: (i) estimating $\mathbf{S}$ assuming $\mathbf{A}$ is fixed, (ii) Inferring the mixing matrix $\mathbf{A}$ assuming $\mathbf{S}$ is fixed. In the simplified GMCA algorithm, the first step boils down to a least-squares estimation of the sources followed by a thresholding as follows :

$$\tilde{\mathbf{\Theta}}_{\mathbf{S}} = \Delta_{\delta}\left(\tilde{\mathbf{A}}^{\dagger}\mathbf{\Theta_X}\right) \ . \tag{59}$$

The next step is a least-squares update of $\mathbf{A}$:

$$\tilde{\mathbf{A}} = \mathbf{\Theta_X}\tilde{\mathbf{\Theta}}_{\mathbf{S}}^{\mathbf{T}}\left(\tilde{\mathbf{\Theta}}_{\mathbf{S}}\tilde{\mathbf{\Theta}}_{\mathbf{S}}^{\mathbf{T}}\right)^{-1} \ . \tag{60}$$

Define $\hat{\mathbf{\Theta}}_{\mathbf{S}} = \tilde{\mathbf{A}}^{\dagger}\mathbf{\Theta}_{\mathbf{X}}$ such that $\tilde{\mathbf{\Theta}}_{\mathbf{S}} = \Delta_{\delta}\left(\hat{\mathbf{\Theta}}_{\mathbf{S}}\right)$ and rewrite the previous equation as follows:

$$\tilde{\mathbf{A}} = \tilde{\mathbf{A}}\hat{\mathbf{\Theta}}_{\mathbf{S}}\Delta_{\delta}\left(\hat{\mathbf{\Theta}}_{\mathbf{S}}\right)^{T}\left(\Delta_{\delta}\left(\hat{\mathbf{\Theta}}_{\mathbf{S}}\right)\Delta_{\delta}\left(\hat{\mathbf{\Theta}}_{\mathbf{S}}\right)^{T}\right)^{-1} . \tag{61}$$

Interestingly, (62) turns out to be a fixed point algorithm with the following stationarity condition :

$$\hat{\mathbf{\Theta}}_{\mathbf{S}}\Delta_{\delta}\left(\hat{\mathbf{\Theta}}_{\mathbf{S}}\right)^{T} = \Delta_{\delta}\left(\hat{\mathbf{\Theta}}_{\mathbf{S}}\right)\Delta_{\delta}\left(\hat{\mathbf{\Theta}}_{\mathbf{S}}\right)^{T} . \tag{62}$$

This fixed point condition constrains $\hat{\mathbf{\Theta}}_{\mathbf{S}}\Delta_{\delta}\left(\hat{\mathbf{\Theta}}_{\mathbf{S}}\right)^{T}$ to be symmetric. In the next section we give a precise probabilistic interpretation to this condition.

5.3.4 Convergence study

In this paragraph, we give some heuristics that enlighten the convergence behavior of the above fast-GMCA algorithm. From a statistical point of view, the sources s_p and s_q are assumed to be random processes. We assume that the entries of $\alpha_p[k]$ and $\alpha_q[k]$ are identically and independently generated from a sparse prior with a heavy-tailed probability density function (*pdf*) which is assumed to be unimodal at zero, even, monotonically increasing for negative values. For instance, any generalized Gaussian distribution verifies these hypotheses. Figure 3 represents the joint *pdf* of two independent sparse sources (on the left) and the joint *pdf* of two mixtures (on the right). We then take the expectation of both sides of (62):

$$\sum_{k\in\Lambda_{\delta}(\hat{\alpha}_q)}\mathbb{E}\{\hat{\alpha}_p[k]\hat{\alpha}_q[k]\} = \sum_{k\in\Lambda_{\delta}(\hat{\alpha}_p)\cap\Lambda_{\delta}(\hat{\alpha}_q)}\mathbb{E}\{\hat{\alpha}_p[k]\hat{\alpha}_q[k]\} , \tag{63}$$

and symmetrically,

$$\sum_{k\in\Lambda_{\delta}(\hat{\alpha}_p)}\mathbb{E}\{\hat{\alpha}_p[k]\hat{\alpha}_q[k]\} = \sum_{k\in\Lambda_{\delta}(\hat{\alpha}_p)\cap\Lambda_{\delta}(\hat{\alpha}_q)}\mathbb{E}\{\hat{\alpha}_p[k]\hat{\alpha}_q[k]\} . \tag{64}$$

Intuitively the sources are correctly separated when the branches of the star-shaped contour plot (see Figure 3 on the left) of the joint *pdf* of the sources are aligned with the axes.
The question is then: *do conditions* (63) *and* (64) *lead to a unique solution ? do acceptable solutions belong to the set of fixed points ?* Note that if the sources are perfectly estimated then $\mathbb{E}\{\Delta_{\delta}\left(\mathbf{\Theta}_{\mathbf{S}}\right)\Delta_{\delta}\left(\mathbf{\Theta}_{\mathbf{S}}\right)^{T}\}$ is diagonal and $\mathbb{E}\{\mathbf{\Theta}_{\mathbf{S}}\Delta_{\delta}\left(\mathbf{\Theta}_{\mathbf{S}}\right)\} = \mathbb{E}\{\Delta_{\delta}\left(\mathbf{\Theta}_{\mathbf{S}}\right)\Delta_{\delta}\left(\mathbf{\Theta}_{\mathbf{S}}\right)\}$. As expected, the set of acceptable solutions (up to scale and permutation) verifies the convergence condition. Let us assume that $\hat{\alpha}_p$ and $\hat{\alpha}_q$ are uncorrelated mixtures of the true sources α_p and α_q. Hard-thresholding then correlates $\hat{\alpha}_p$ and $\Delta_{\delta}(\hat{\alpha}_q)$ (respectively $\hat{\alpha_q}$ and $\Delta_{\delta}(\hat{\alpha}_p)$) unless the joint *pdf* of the estimated sources α_p and α_q has the

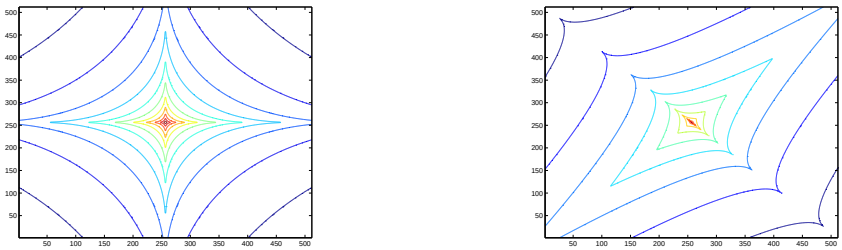

Fig. 3. Contour plots of a simulated joint *pdf* of 2 independent sources generated from a generalized Gaussian law $f(x) \propto \exp(-\mu|x|^{0.5})$. **Left :** joint *pdf* of the original independent sources. **Right :** joint *pdf* of 2 mixtures.

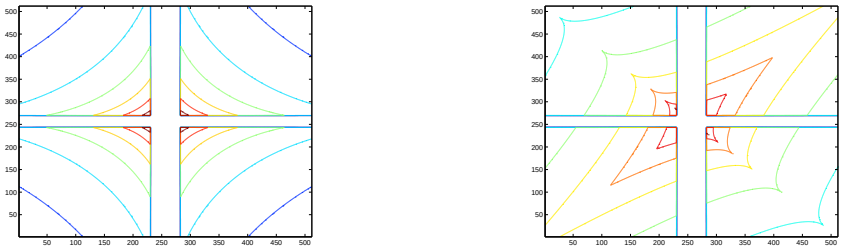

Fig. 4. Contour plots a simulated joint *pdf* of 2 independent sources generated from a generalized Gaussian law that have been hard-thresholded. **Left :** joint *pdf* of the original independent sources that have been hard-thresholded. **Right :** joint *pdf* of 2 mixtures of the hard-thresholded sources.

same symmetries as the thresholding operator (this property has also been outlined in (26)). Figure 4 gives a rather good empirical point of view of the previous remark. On the left, Figure 4 depicts the joint *pdf* of two unmixed sources that have been hard-thresholded. Note that whatever thresholds we apply, the thresholded sources are still decorrelated as their joint *pdf* has the same symmetries as the thresholding operator. On the contrary, on the right of Figure 4, the hard-thresholding process further correlates the two mixtures. For a fixed δ, several fixed points lead to decorrelated coefficient vectors $\hat{\alpha}_p$ and $\hat{\alpha}_q$. Figure 4 provides a good intuition: for fixed δ the set of fixed points is divided into two different categories: (i) those which *depend* on the value of δ (plot on the right) and (ii) those that are valid fixed points for all values of δ (plot on the left of Figure 4). The latter solutions lead to acceptable sources up to scale and permutation. Remark that those conditions must hold for every threshold $\delta \geq \delta^\star$, where $\delta^\star$ is the minimum scalar δ such that the sources s_p and s_q are δ-disjoint. As fast-GMCA involves a decreasing thresholding scheme, the final fixed points are stable if they verify the convergence conditions (63) and (64) for all δ. To conclude, if the fast-GMCA algorithm converges, it should converge to the true sources up to scale and permutation.

We finally note that noise is naturally handled in the accelerated GMCA as for the original version. For instance, in presence of noise, the MCA used in the first step to get the sparse decomposition of the observations, is stopped

at typically $3 - 4\sigma_{\mathbf{N}}$. This strategy will be supported by the experiments of Section 5.6.

5.4 When the number of sources is unknown

In blind source separation, the number of sources is assumed to be a fixed known parameter of the problem. In practical situations, the number of sources is often rarely known and has to be estimated. In an ideal theoretical setting, the number of sources is the dimension of the subspace of $\mathbb{R}^m$ (recall that m is the number of observations or channels) in which the data lies. A mis-estimation of the number of sources n may entail several difficulties:

- Under-estimation : in the GMCA algorithm, under-estimating the number of sources will clearly lead to solutions that are made of linear combinations of "true" sources. The solution may then be sub-optimal with respect to the sparsity of the estimated sources.
- Over-estimation : in case of over -estimation, the GMCA algorithm may have to cope with a mixing matrix estimate that has not a full column-rank. The optimization problem at hand can be ill-conditioned.

Henceforth, estimating the number of sources is a crucial and strenuous issue. To our knowledge, only a few work have already focused on the estimation of the number of sources n. Recently, the author in (89) approached the problem using the minimum description length.In this paper, we introduce a sparsity-based method to estimate n within the GMCA framework.
It is possible, as in (89), to use classical model selection criteria in the GMCA algorithm. Such criteria, including AIC (90), BIC (91), would provide a balance between the complexity of the model (here the number of sources) and its ability to faithfully represent the data. It would amount to add a penalty term in (40). This penalty term would merely prevent a high number of sources.

In the sparse BSS framework, we propose an alternative approach. Indeed, for a fixed number of sources $p < n$, the sparse BSS problem amounts to solving the following optimization task :

$$\min_{\mathbf{A},\alpha \,\big|\, \mathrm{ColDim}(\mathbf{A})=p} \|\alpha\|_1 \;\text{ s.t. }\; \|\mathbf{X} - \mathbf{A}\alpha\mathbf{\Phi}\|_F < \epsilon \;. \tag{65}$$

where $\mathrm{ColDim}\,(\mathbf{A})$ is the number of columns of the matrix $\mathbf{A}$. The general algorithm we would like to deal with is then the following :

$$\min_{p} \left\{ \min_{\mathbf{A},\alpha \,\big|\, \mathrm{ColDim}(\mathbf{A})=p} \|\alpha\|_1 \;\text{ s.t. }\; \|\mathbf{X} - \mathbf{A}\alpha\mathbf{\Phi}\|_F < \epsilon \right\} \;. \tag{66}$$

Let us write $\mathcal{P}_{p,\epsilon}$ the problem in Equation (65). Interestingly, if $p < n$, there exists a minimal value $\epsilon^\star(p)$ such that if $\epsilon < \epsilon^\star(p)$, problem $\mathcal{P}_{p,\epsilon}$ has no solution. For a fixed $p < n$, this minimal value $\epsilon^\star(p)$ is obtained by approximating the data $\mathbf{X}$ with its largest p singular vectors.
Furthermore, in the noiseless case, for $p < n$, $\epsilon^\star(p)$ is always strictly positive as the data lies in a subspace the whose dimension is exactly n. Then, when $p = n$, the problem $\mathcal{P}_{n,\epsilon}$ has at least one solution for $\epsilon = \epsilon^\star(n) = 0$. Then, devising a joint estimation scheme for the mixing matrix $\mathbf{A}$, the sources $\mathbf{S}$ and the number of sources n is possible via a constructive approach. Indeed, we propose to look for the solutions of $\mathcal{P}_{p,\epsilon}$ for increasing values of $p \geq 1$ and varying values of ϵ. As the GMCA algorithm is likely to provide the solution of $\mathcal{P}_{p,\epsilon^\star(p)}$ for a fixed p, we propose the following GMCA-based algorithm :

While $\|\mathbf{X} - \mathbf{A}\alpha\mathbf{\Phi}\|_F > \epsilon^\star(n)$ and $p \leq m$:

1- Increase p by adding a new column to $\mathbf{A}$ - this step is described below.

2- Solve $\mathcal{P}_{p,\epsilon^\star(p)}$ using the GMCA algorithm for a fixed p :

$$\min_{\mathbf{A},\alpha \left| \mathsf{ColDim}(\mathbf{A})=p \right.} \|\alpha\|_1 \text{ s.t. } \|\mathbf{X} - \mathbf{A}\alpha\mathbf{\Phi}\|_F < \epsilon^\star(p) \ .$$

The role of the GMCA algorithm : The above algorithm then strives to find a particular path described by a sequence $\{p_i, \epsilon_i\}_i$ of solutions to $\mathcal{P}_{p_i,\epsilon_i}$. Ideally, an optimal scheme would provide the sequence $\{i, \epsilon^\star(i)\}_{i=1,\cdots,n}$ of solutions to $\mathcal{P}_{i,\epsilon^\star(i)}$ thus leading to the optimal value $\epsilon^\star(n) = 0$ when $i = n$. Nevertheless, this sequence is hard to obtain in practice : the optimal sequence $\{i, \epsilon^\star(i)\}_{i=1,\cdots,n}$ is unknown *a priori*. Hopefully, for a fixed p, the way the threshold decreases and stops in the GMCA algorithm (Step 2 of the above algorithm) should enable GMCA to provide a solution close to $\mathcal{P}_{p,\epsilon^\star(p)}$. Indeed, in the GMCA framework, there is a bijective map between a value of ϵ in Equation (65) and the threshold δ used in the GMCA algorithm such that both formulations share the same solution. Obviously, when $\epsilon \to 0$ then $\delta \to 0$. Then, in practice, for a fixed value of p, managing the threshold such that it tends to 0 in the GMCA algorithm should lead to a solution close to $\mathcal{P}_{p,\epsilon^\star(p)}$.
In the noiseless case, Step 2 of the above algorithm then amounts to running a whole GMCA estimation of $\mathbf{A}$ and $\mathbf{S} = \alpha\mathbf{\Phi}$ for a fixed p with a final threshold $\delta_{\min} = 0$.
Let us remark that the sequence $\{i, \epsilon^\star(i)\}_{i=1,\cdots,n}$ can be estimated in advance. Indeed, for a fixed number of components $p = i$, the optimal approximation error is given by the projection on the subspace of dimension p spanned by the p singular vectors related to the p highest singular values of $\mathbf{X}$. A pre-

processing step would require to compute the singular value decomposition of $\mathbf{X}$ to estimate the optimal sequence $\{i, \epsilon^\star(i)\}_{i=1,\cdots,n}$. In practical situations, the use of GMCA avoids this pre-processing step.

Increasing iteratively the number of components : In the aforementioned algorithm, Step 1 amounts to adding a column vector to the current mixing matrix $\mathbf{A}$. The most simple choice would amount to choose this vector at random. Wiser choices can also be made based on additional priori information :

- **Decorrelation :** if the mixing matrix is assumed to be orthogonal, the new column vector can be chosen as being orthogonal to the subspace spanned by the columns of $\mathbf{A}$ with $\text{ColDim}(\mathbf{A}) = p - 1$.
- **Known spectra :** if a set of spectra are known *a priori*, the new column can be chosen amongst the set of unused spectra. The new spectrum can be chosen based on its coherence with the residual. Let $\mathcal{A}$ denote a set of spectra $\{\eta_l \in \mathcal{A}\}_{l=1,\cdots,\text{Card}(\mathcal{A})}$ and let note $\mathcal{A}_c$ the set of unused spectra (*i.e.* spectra that have not been chosen previously), then the p^{th} column of $\mathbf{A}$ is chosen such that :

$$\eta_{l^\star} = \arg\max_{\eta_l \in \mathcal{A}_c} \left| \sum_{k=1}^{t} \frac{1}{\|\eta_l\|_{\ell_2}^2} \eta_l^T \left[\mathbf{X} - \mathbf{AS}\right]^k \right| , \tag{67}$$

  where $[\mathbf{X} - \mathbf{AS}]^k$ is the k^{th} column of $\mathbf{X} - \mathbf{AS}$.

Any other prior information can be taken into account which will guide the choice of a new column vector of $\mathbf{A}$.

The noisy case : In the noisy case, the parameter ϵ^2 can be interpreted as a bound on noise (for bounded noise such as the case of Gaussian white noise with covariance matrix $\sigma_{\mathbf{N}}^2\mathbf{I}$). In the second probabilistic case, the noise is known to be bounded above and below by $\pm\pi\sigma_{\mathbf{N}}$ with probability higher than $1 - \exp(-\pi^2/2)$. In practice, in Step 2 of the above algorithm, the final threshold of the GMCA algorithm is chosen as $\delta_{\min} \simeq 3\sigma_{\mathbf{N}}$. The choice $\pi = 3$ then guarantees the noise to be bounded with probability higher than 0.98.

A simple experiment

In this experiment, the data are assumed to be the linear combination of n sources as stated by the classical instantaneous mixture model. The entries of $\mathbf{S}$ have been independently drawn from a Laplacian probability density with

scale parameter $\mu = 1$ ($\mathbf{\Phi}$ is chosen as the Dirac basis). The entries of the mixing matrix are independently drawn form a zero-mean Gaussian distribution of unit-variance. The data are not contaminated by noise.
This experiment will focus on comparing the classical PCA (the popular subspace selection method), and the GMCA algorithm assuming n is unknown. In the absence of noise contaminating the data, only the n highest eigenvalues provided by the PCA, which coincide with the Frobenius norm of each product $\{a^i s_i\}_{i=1\cdots,n}$, are non-zero. PCA therefore provides the true number of sources. In Figure 5, the aforementioned GMCA algorithm has been applied to the same data in order to estimate the number of sources. In this experiment, the number of channels is $m = 64$. Each observation has $t = 256$ samples. The number of sources n varies from 2 to 20. Each point has been computed from 25 trials. Figure 5 depicts the mean value of the number of sources estimated by GMCA. For each of the 25 trials, GMCA provides exactly the true number of sources.

In Figure 6, we compare the performances of PCA and GMCA in recov-

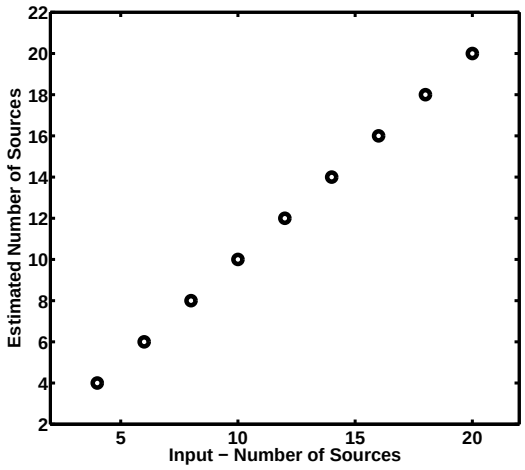

Fig. 5. **Estimating the number of sources with GMCA- Abscissa :** True number of sources. **Ordinate :** Estimated number of sources with GMCA. Each point is the mean number of sources computed from 25 trials. For each point, the estimation variance is zero.

ering the true input sources. In this experiment, the number of channels is $m = 128$. Each channel has $t = 2048$ samples. The top panel of Figure 6 shows the mean recovery SNR in dB of the estimated sources. Clearly, the GMCA provides sources that are closer to the true sources than PCA. Let us define the following ℓ_1-norm based criterion:

$$\mathcal{C}_{\ell_1} = \frac{\sum_{i=1}^{n} \left\| a^i s_i - \tilde{a}^i \tilde{s}_i \right\|_1}{\sum_{i=1}^{n} \left\| a^i s_i \right\|_1} \ , \tag{68}$$

where symbol $\sim$ means estimated parameters. $\mathcal{C}_{\ell_1}$ provides a sparsity-based criterion that quantifies the deviation between the estimated sources and the true sparsest sources. The panel at the bottom of Figure 6 shows the evolution of $\mathcal{C}_{\ell_1}$ when the number of sources varies. As expected, the GMCA-based algorithm also provides the sparsest solutions.

These preliminary examples point out that GMCA is able to find the true

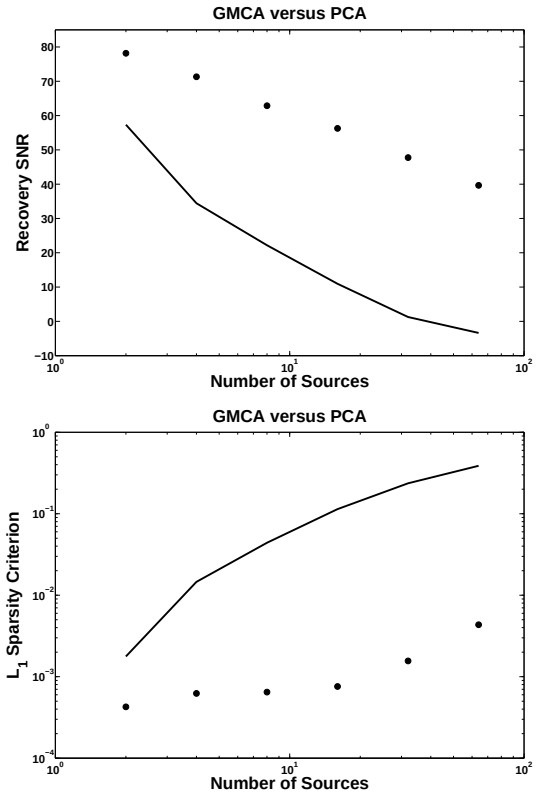

Fig. 6. **GMCA (dots) versus PCA (solid) - Abscissa :** Input number of sources. **Ordinate - Top :** Recovery SNR in dB **Bottom :** ℓ_1 sparsity criterion. Each point is an average value over 25 trials.

dimension of the subspace in which the data lies (*i.e.* the true number of sources). Furthermore, GMCA provides far sparser solutions than PCA with much smaller recovery errors. Further work is needed to better characterize the behavior of GMCA when the number of sources is unknown. This is clearly one perspective to consider in a future work.

5.5 Variations on sparsity and independence

Up to now, we considered the data $\mathbf{X}$ as a collection of m channels each of which having t entries or samples. Considering instead $\mathbf{X}$ as a collection of t

signals having m entries (the columns of the matrix $\mathbf{X}$) leads to an interesting point of view. In this paragraph, we assume that the data $\mathbf{X}$ has already been decomposed in the *spatial* dictionary $\mathbf{\Phi}$ such as in Section 5.3. We then handle the coefficients of $\mathbf{X}$ in $\mathbf{\Phi}$ defined as follows :

$$\mathbf{X} = \mathbf{\Theta_X \Phi} \ . \tag{69}$$

For the sake of simplicity, we will assume that $\mathbf{\Phi}$ is a nice orthonormal matrix; $\mathbf{\Theta_X}$ then has the same dimension as $\mathbf{X}$. We will also assume that $m = n$ and that the mixing matrix $\mathbf{A}$ is invertible. Similarly, the sources are represented in $\mathbf{\Phi}$ via their decomposition coefficients $\mathbf{\Theta_S}$. We assume that each entry of $\mathbf{\Theta_S}$ is random and generated from a Laplacian distribution with scale parameter μ. The entries of $\mathbf{\Theta_S}$ are mutually independent. Recalling that the $\{i,j\}^{th}$ entry of $\mathbf{\Theta_S}$ is written α_{ij}, the probabilistic model is defined as follows :

$$\forall i = 1, \cdots, n : j = 1, \cdots, t; \quad f\left(\alpha_{ij}\right) \propto \exp\left(-\mu|\alpha_{ij}|\right) \ . \tag{70}$$

The matrix $\mathbf{\Theta_S}$ can be viewed has the concatenation of t $n \times 1$ column vectors $\{\theta_S^k\}_{k=1,\cdots,T}$ such that:

$$\forall k = 1, \cdots, t; i = 1, \cdots, n; \quad \theta_S^k[i] = \alpha_{ij} \ . \tag{71}$$

Clearly, the set of vectors $\{\theta_S^k\}_{k=1,\cdots,T}$ are mutually independent and their individual probability function is as follows :

$$f_{\mathbf{S}}(\theta_S^k) = \prod_{i=1}^{n} f_{\mathbf{S}}(\theta_S^k[i]) \propto \exp\left(-\mu\|\theta_S^k\|_1\right) \ . \tag{72}$$

The *noiseless* sparse BSS problem then amounts to looking for the matrix $\mathbf{B} = \mathbf{A}^{-1}$ that minimizes the sparsity of the estimated sources. From the point of view of optimization, the problem can be rewritten as follows :

$$\min_{\mathbf{A},\{\theta_S^k\}} \sum_{k=1}^{t} \left\|\theta_S^k\right\|_{\ell_1} \text{ s.t. } \mathbf{\Theta_X} = \mathbf{A\Theta_S}. \tag{73}$$

Assuming $\mathbf{A}$ is fixed, the problem above is equivalent to recovering the sparse decomposition of each column of $\mathbf{\Theta_X}$ **separately** in the *dictionary* $\mathbf{A}$. In the general case, the mixing matrix $\mathbf{A}$ is unknown. The problem in Equation (73) is then equivalent to seeking the *basis* (not necessarily orthogonal) $\mathbf{A}$ in which the columns of $\mathbf{\Theta_X}$ are **jointly** the sparsest. This problem is then quite similar to the search for the "best sparsifying basis" described in (92).
In this framework, the sparse BSS issue is equivalent solving the "best sparsifying basis" problem for an ensemble of vectors. In the next paragraphs we go further and exhibit close links between different problems as summarized in the diagram:

$$\text{Best Sparsifying/Unconditional Bases} \ \leftrightarrow \ \text{Sparse BSS} \ \leftrightarrow \ \text{ICA}$$

In the probabilistic framework described in (72), the set of vectors $\{\theta_S^k\}_{k=1,\cdots,T}$ belongs with high probability (if $C \gg 1/\mu$) to the ℓ_1-ball $\mathbf{\Omega} = \{\theta \Big| \|\theta\|_1 \leq C\}$. Furthermore, each column vector θ_S^k is, by definition, the transformed version of the original source column vectors θ_S^k :

$$\forall k = 1, \cdots, t; \quad \theta_X^k = \mathbf{A}\theta_S^k \tag{74}$$

We further assume that $\mathbf{A}$ has columns with unit ℓ_2-norm. It follows that the set of mixed vectors $\{\theta_X^k\}_{k=1,\cdots,T}$ belongs, with high probability, to $\mathbf{A\Omega}$ the image of the ℓ_1-ball $\mathbf{\Omega}$ by $\mathbf{A}$.
In this particular framework, we can find very close connections with the work of Donoho in (43) in a different context. Inspired by this work, we can transpose exactly the same results to our framework. Indeed, the vectors $\{\theta_X^k\}_{k=1,\cdots,T}$ must have a kind of unique "unconditional basis" (see (43)) which turns out to be A^{-1}. Conversely, looking for the sparsest representation of the set of vectors $\{\theta_X^k\}_{k=1,\cdots,T}$ (with respect to the ℓ_1 metric) solves the sparse BSS issue. Inspired by (43), the following proposition gives mild conditions proving that the sparsest solution provides the sparse BSS solution.

Proposition 3 *Assume that the sources, in the sparse domain,* $\mathbf{\Theta_S}$ *have entries independently and identically distributed from a Laplacian density. Assume that the mixing matrix A is invertible, has columns with unit* ℓ_2 *norm. Then, the norm* $\|\cdot\|_{1,1}$ *is a "contrast" function :*

$$\mathbb{E}\left\{\|\mathbf{\Theta_S}\|_1\right\} \leq \mathbb{E}\left\{\|\mathbf{A\Theta_S}\|_1\right\} \ . \tag{75}$$

The proof is inspired by that of Lemma 4 in (43).
The latter viewpoint of the sparse BSS problem yields several conclusions:

- Contrast function : The ℓ_1-norm is a contrast function. Indeed, looking for the sparsest solutions provides the solutions of the BSS problem.
- Unconditional basis : Looking for a demixing matrix can be equivalently done by seeking the "unconditional basis" of the set of vectors $\{\theta_X^k\}_{k=1,\cdots,T}$.

Note that in harmonic analysis, the search for unconditional bases is motivated by their ability to provide the so-called "diagonal" processes. In the sparse BSS framework, the "diagonality" property is no more than the independence of the entries of the column vectors $\{\theta_S^k\}_{k=1,\cdots,T}$. This remark clearly stresses the interplay between apparently different concepts, namely sparse BSS, unconditional bases and ICA. Interestingly, Meyer in (93) has already pointed out the intuitive link between ICA and unconditional basis.

From the ICA viewpoint, the Laplacian prior may be exploited via a Maximum Likelihood approach. Estimating the sources in the ML framework amounts to solving the following optimization problem :

$$\min_{\mathbf{B}} \|\mathbf{B}\mathbf{\Theta}_\mathbf{X}\|_1 - \log|\det(\mathbf{B})| \ . \tag{76}$$

In the noiseless case, the equality condition $\mathbf{\Theta}_\mathbf{X} = \mathbf{A}\mathbf{\Theta}_\mathbf{S}$ enables to recast the above problem as follows :

$$\min_{\mathbf{A},\mathbf{\Theta}_\mathbf{S}} \|\mathbf{\Theta}_\mathbf{S}\|_1 \ , \text{ s.t. } \mathbf{\Theta}_\mathbf{X} = \mathbf{A}\mathbf{\Theta}_\mathbf{S} \tag{77}$$

which is valid when $\log|\det(\mathbf{A})|$ is constant (for instance in the orthogonal case). Then, the above problem is directly equivalent to the sparse BSS problem described in Equation (73). As a consequence, sparse ICA is equivalent to sparse BSS.

5.6 Results

The sparser, the better

Up to now we used to claim that sparsity and morphological diversity are the clue for good separation results. The role of morphological diversity is twofold:

- **Separability :** the sparser the sources in the dictionary $\mathbf{\Phi}$ (redundant or not), the more "separable" they are. As we noticed earlier, sources with different morphologies are diversely sparse (*i.e.* they have δ-disjoint supports in $\mathbf{\Phi}$ with a "small" δ). The use of a redundant $\mathbf{\Phi}$ is thus motivated by the grail of sparsity for a wide class of signals for which sparsity means separability.
- **Robustness to noise or model imperfections :** the sparser the sources, the less dramatic the noise. In fact, sparse sources are concentrated on very few significant coefficients in the sparse domain for which additive noise is a slight perturbation. As a sparsity-based method, GMCA should be less sensitive to noise.

Furthermore, from a signal processing point of view, dealing with highly sparse signals leads to easier and more robust models. To illustrate those points, let us consider $n = 2$ unidimensional sources with $t = 1024$ samples. These sources are the *Bump* and *HeaviSine* signals available in the WaveLab toolbox - see (94). The first column of Figure 7 shows the two synthetic sources. The sources are randomly mixed, and a Gaussian noise with variance corresponding to SNR=19dB is added so as to provide $m = 2$ observations portrayed in the

second column of Figure 7. We assumed that MCA preserves linearity for such sources and mixtures (see our choice of the dictionary later on). The mixing matrix is assumed to be unknown. The third and fourth columns of Figure 7 depict the GMCA estimates computed with respectively (i) a single orthonormal discrete wavelet transform (DWT) and (ii) a union of DCT and DWT. Visually, GMCA performs quite well in both cases.

Figure 8 gives the value of the mixing matrix criterion $\mathcal{C}_{\mathbf{A}}$ (defined in section

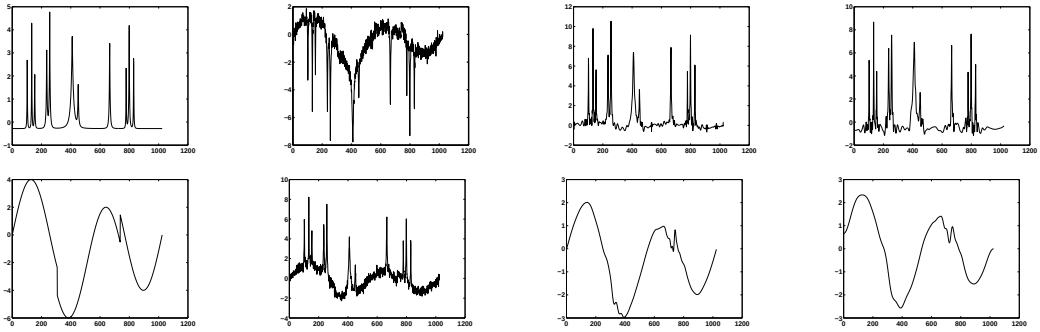

Fig. 7. **The sparser the better - first column:** the original sources. **Second column:** mixtures with additive Gaussian noise (SNR = 19dB). **Third column:** sources estimated with GMCA using a single Discrete Orthogonal Wavelet Transform (DWT). **Fourth column:** Sources estimated with GMCA using a redundant dictionary made of the union of a DCT and a DWT.

5.2) as the SNR increases. In Figure 8, the *dashed* line corresponds to the behavior of GMCA in a single DWT; the *solid* line depicts the results obtained using GMCA when $\mathbf{\Phi}$ is the union of the DWT and the DCT. On the one hand, GMCA gives satisfactory results as $\mathcal{C}_{\mathbf{A}}$ is rather low for both experiments. On the other hand, the values of $\mathcal{C}_{\mathbf{A}}$ provided by GMCA in the MCA-domain are approximately 5 times better than those given by GMCA using a unique DWT. This simple toy experiment clearly confirms the benefits of sparsity for blind source separation. Furthermore it underlines the effectiveness of "very" sparse representations provided by non-linear decompositions in overcomplete dictionaries. This is an occurrence of what D.L. Donoho calls the "blessing of dimensionality" (95).

GMCA is able to provide the sparsest solution

In this paragraph, we have run a simple noiseless experiment. The data $\mathbf{X}$ consists of 4 mixtures (Figure 10) each of which is the linear combination of 4 sources (Figure 9). The mixing matrix has been chosen at random. The GMCA algorithm has been performed in the biorthogonal wavelet domain; see (71). The estimated sources are shown in Figure 11. These results were obtained using the GMCALab toolbox (96).

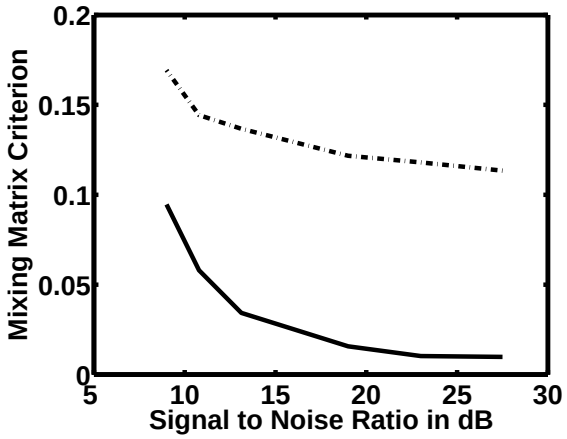

Fig. 8. **The sparser the better :** behavior of the mixing matrix criterion when the noise variance increases for DWT-GMCA (Dashed line) and (DWT+DCT)-GMCA (Solid line).

We previously emphasized on GMCA as being able to provide the sparsest sources in the sense advocated by the sparse BSS framework. Figure 12 provides the evolution of the sparsity divergence $\|\tilde{\mathbf{S}}\|_1 - \|\mathbf{S}\|_1$ along the 500 GMCA iterations. Clearly, the GMCA algorithm tends to estimate sources with increasing sparsity. Furthermore, the GMCA solution has the same sparsity (with respect to the sparsity divergence) as the true sources. This simple experiment then points out that GMCA is able to recover the solution having the correct sparsity level.

Fig. 9. The 256×256 source images.

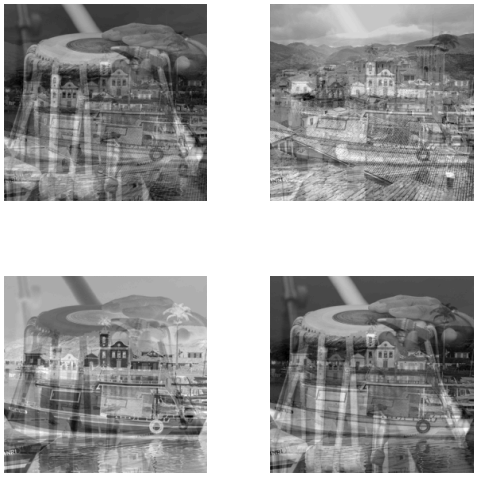

Fig. 10. The 256×256 noiseless mixtures.

Fig. 11. The sources estimated using GMCA.

Dealing with noise

The last paragraph emphasized on sparsity as the key for very efficient source separation methods. In this section, we will compare several BSS techniques with GMCA in an image separation context. We chose 3 different reference BSS methods:

- JADE : the well-known ICA (Independent Component Analysis) based on fourth-order statistics (see (16)).
- Relative Newton Algorithm : the seminal sparsity-based BSS technique of

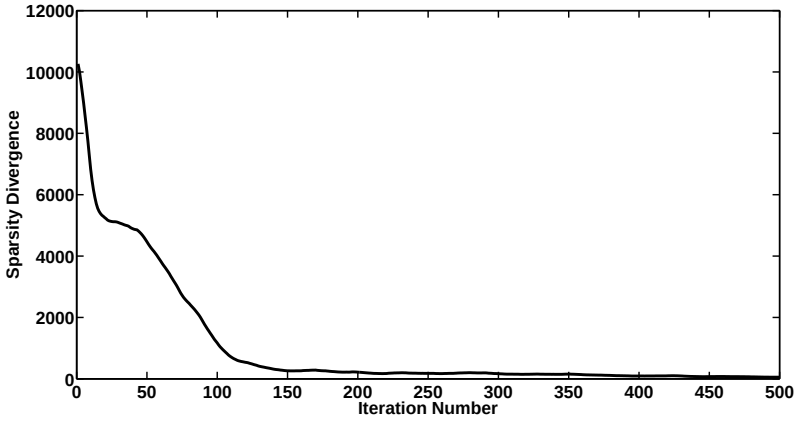

Fig. 12. **GMCA provides the sparsest solution - Abscissa :** Iteration number. **Ordinate :** Sparsity divergence $\|\tilde{\mathbf{S}}\|_1 - \|\mathbf{S}\|_1$.

Zibulevsky (29) we already reviewed. In the experiments reported hereafter, we used the Relative Newton Algorithm (RNA) on the data transformed by a basic orthogonal 2D wavelet transform (2D-DWT).

- EFICA : this separation method improves the FastICA algorithm for sources following generalized Gaussian distributions (leptokurtic marginals with heavy tails). We also applied EFICA on data transformed by a 2D-DWT where the leptokurticity assumption on the source marginal statistics is valid.

Figure 13 shows the original sources (top pictures) and the 2 mixtures (bottom pictures). The original sources s_1 and s_2 have a unit variance. The matrix $\mathbf{A}$ that mixes the sources is such that $x_1 = 0.25s_1 + 0.5s_2 + n_1$ and $x_2 = -0.75s_1 + 0.5s_2 + n_2$ where n_1 and n_2 are Gaussian noise vectors (with decorrelated samples) such that SNR=10dB. The noise covariance matrix $\mathbf{\Sigma_N}$ is diagonal.

In section 5.6 we claimed that a sparsity-based algorithm would lead to more robustness to noise. The comparisons we carry out here are twofold: (i) we evaluate the separation quality in terms of the correlation of the original and estimated sources as the noise variance varies; (ii) as the estimated sources are also perturbed by noise, correlation coefficients are not always very sensitive to separation errors so that we also assess the performances of each method by computing the mixing matrix criterion $\mathcal{C}_{\mathbf{A}}$. The GMCA algorithm was applied using a dictionary consisting of the union of a Fast Curvelet Transform (available online - see (74; 97)) and a Local Discrete Cosine Transform (LDCT). The union of the curvelet transform and LDCT are often well suited to a wide class of "natural" images.

Figure 14 portrays the evolution of the correlation coefficient of source 1 (left picture) and source 2 (right picture) as a function of the SNR. At first glance, GMCA, RNA and EFICA are very robust to noise as they give correlation co-

Fig. 13. **Top :** the 256 × 256 source images. **Bottom :** two different mixtures. Gaussian noise is added such that the SNR is equal to 10dB.

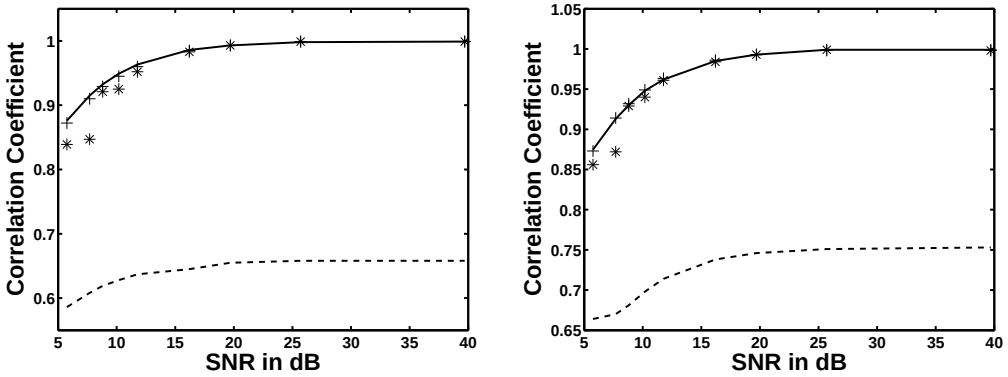

Fig. 14. Evolution of the correlation coefficient between original and estimated sources as the noise variance varies: *solid* line : GMCA, *dashed* line: JADE, *(⋆) :* EFICA, *(+) :* RNA. **Abscissa :** SNR in dB. **Ordinate :** correlation coefficients.

efficients close to the optimal value 1. On these images, JADE behaves rather badly. It might be due to the correlation between these two sources. For higher noise levels (SNR lower than 10dB), EFICA tends to perform slightly worse than GMCA and RNA. As we noted earlier, in our experiments, a mixing matrix-based criterion turns out to be more sensitive to separation errors and then better discriminates between the methods. Figure 15 depicts the behavior of the mixing matrix criterion as the SNR increases. Recall that the correlation coefficient was not able to discriminate between GMCA and RNA. The

mixing matrix criterion clearly reveals the differences between these methods. First, it confirms the dramatic behavior of JADE on that set of mixtures. Secondly, RNA and EFICA behave rather similarly. Thirdly, GMCA seems to provide far better results with mixing matrix criterion values that are up to 10 times better than JADE and approximately 2 times better than with RNA or EFICA.

To summarize, the findings of this experiment confirm the key role of sparsity in blind source separation:

- **Sparsity brings better results :** remark that, amongst the methods we used, only JADE is not a sparsity-based separation algorithm. Whatever the method, separating in a sparse representation enhances the separation quality : RNA, EFICA and GMCA clearly outperform JADE.
- **GMCA takes better advantage of overcompleteness and morphological diversity:** RNA, EFICA and GMCA provide better separation results with the benefit of sparsity. Nonetheless, GMCA takes better advantage of overcomplete sparse representations than RNA and EFICA.

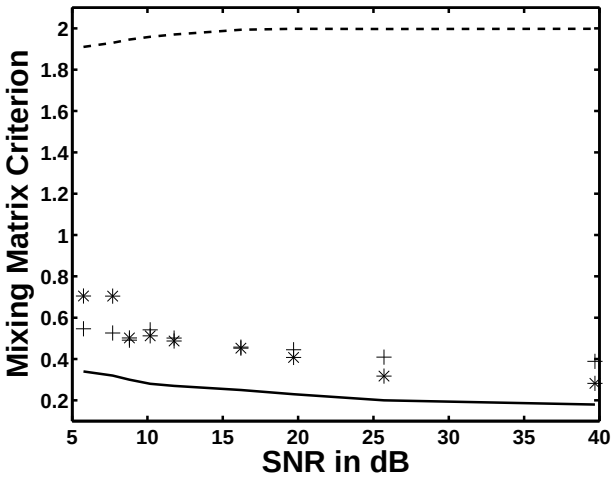

Fig. 15. Evolution of the mixing matrix criterion Δ_A as the noise variance varies: *solid* line : GMCA, *dashed* line : JADE, *(⋆) :* EFICA, *(+) :* RNA. **Abscissa :** SNR in dB. **Ordinate :** mixing matrix criterion value.

Higher dimension problems and computational cost

In this section, we propose to analyze how GMCA behaves when the dimension of the problem increases. Indeed, for a fixed number of samples t, it would be more difficult to separate mixtures with a high number of sources n. In the following experiment, GMCA is applied on data that are random mixtures of $n = 2$ to 15 sources. The number of mixtures m is set to be equal to the number of sources : $m = n$. The sources are selected from a set of 15 images (of size 128×128 pixels). These sources are depicted in Figure 16. GMCA was applied

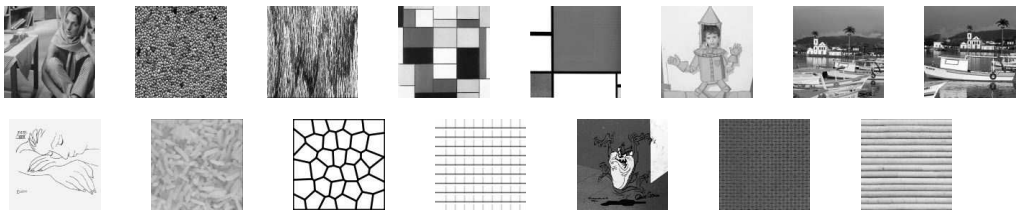

Fig. 16. The set of 15 sources used to analyze how GMCA scales when the number of sources increases.

using the Fast Curvelet transform (74). Hereafter, we analyze the convergence of GMCA in terms of the mixing matrix criterion $\mathcal{C}_{\mathbf{A}}$. This criterion is normalized as follows $\bar{\mathcal{C}}_{\mathbf{A}} = \frac{\mathcal{C}_{\mathbf{A}}}{n^2}$ to be independent of the number of sources n. The plot on the left of Figure 17 shows how GMCA behaves when the number of iterations $I_{\max}$ varies from 2 to 1000. Whatever the number of sources, the normalized mixing matrix criterion drops when the number of iterations is higher than 50. When $I_{\max} > 100$, the GMCA algorithm tends to stabilize. Then, increasing the number of iterations does not lead to a substantial separation enhancement. When the dimension of the problem increases, the normalized mixing matrix criterion at convergence gets slightly larger ($I_{\max} > 100$). As expected, for a fixed number of samples t, the separation task is likely to be more difficult when the number of sources n increases. Fortunately, GMCA still provides good separation results with low mixing matrix criterion (lower than 0.025) values up to $n = 15$ sources.
The plot on the right of Figure 17 illustrates how the computational cost [6] of GMCA scales when the number of sources n varies. Recall that the fast GMCA algorithm is divided into two steps : i) sparsifying the data and compute $\mathbf{\Theta}_{\mathbf{X}}$, ii) estimating the mixing matrix $\mathbf{A}$ and $\mathbf{\Theta}_{\mathbf{S}}$. This plot shows that the computational burden obviously increases when the number of sources n grows. Let us point out that, when $m = n$, the computational burden of step i) is proportional to the number of sources n and independent of the number of iterations $I_{\max}$. Then, for high $I_{\max}$ values, the computational cost of GMCA tends to be proportional to the number of iterations $I_{\max}$.

6 Dealing with Hyperspectral Data

6.1 Specificity of hyperspectral data

Considering the objective function in the minimization problem (27) from a Bayesian perspective, the ℓ_1 penalty terms imposing sparsity are easily interpreted as coming from Laplacian prior distributions on the components s_k and problem (27) is akin to a Maximum A Posteriori estimation of the model

[6] The experiments were run with IDL on a PowerMac G5 - 2Ghz computer.

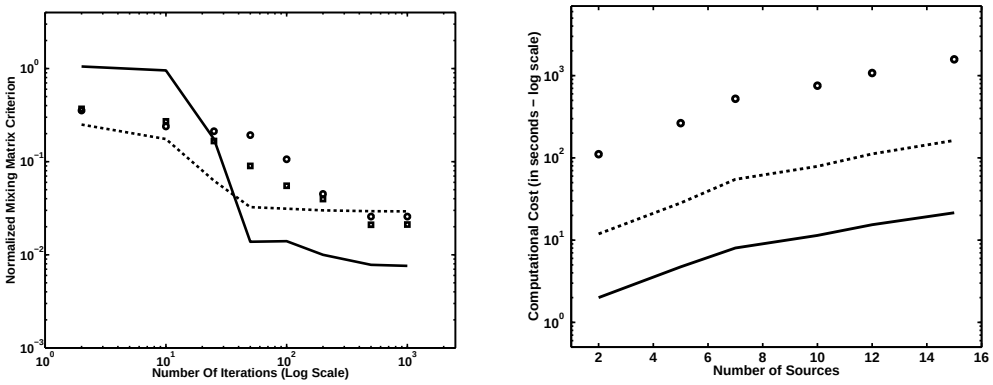

Fig. 17. **Left :** Evolution of the normalized mixing matrix criterion when the number of GMCA iterations $I_{\max}$ increases. **Abscissa**: Number of iterations. **Ordinate :** Normalized mixing matrix criterion. The number of sources varies as follows *solid* line : $n = 2$, *dashed* line : $n = 5$, *(□)* : $n = 10$, *(∘)* : $n = 15$. **Right :**. Behavior of the computational cost when the number of sources increases. **Abscissa**: Number of sources. **Ordinate :** Computational cost in seconds. The number of iterations varies as follows *solid* line : $I_{\max} = 10$, *dashed* line : $I_{\max} = 100$, *(∘)* : $I_{\max} = 1000$.

parameters **A** and **S**. Interestingly, there is a striking asymmetry in the treatment of **A** and **S** and this is in fact a common feature of the great majority of BSS methods. Invoking a uniform *improper* prior distribution for the spectral parameters **A** is standard practice. On the one hand, this unbalanced treatment may not seem so unfair when **A** and **S** actually do have very different roles in the model and very different sizes. As mentioned earlier, **A** is often simply seen as a mixing matrix of small and fixed size while each row s_i of the source matrix **S**, is usually seen as a collection of t samples from a process in time or pixels in an image, which can grow very much larger than the number of channels m as more data is collected. On the other hand, there are applications in which one deals with data from instruments with a very large number of channels which are well organized according to some physically meaningful index. A typical example is hyperspectral data where images are collected in a large number of, what is more, contiguous regions of the electromagnetic spectrum. It then makes sense to consider the continuity, the regularity, etc. of some physical property from one channel to its neighbor. For instance the spectral signatures of the objects in the scene may be known *a priori* to have a sparse representation in some specified possibly redundant dictionary of *spectral* waveforms.
In what follows, the term *hyperspectral* is used generically to identify data with the following specific properties regardless of other definitions or models living in other scientific communities :

(1) **High dimensionality :** The number of channels m in common hyperspectral imaging devices can be greater than a hundred. Consequently, problems involving hyperspectral data often have very high dimensions.

(2) **Contiguity :** The large number of channels in the instrument achieve a *regular / uniform* sampling of some additional and meaningful physical index (wavelength, space, time). We refer to this added dimension as the *spectral* dimension.

(3) **Morphospectral Coherence :** Hyperspectral data is assumed to be structured *a priori* according to the linear mixture model given in equation (3).

We describe next an extension of the GMCA algorithm for hyperspectral data processing when it is known *a priori* that the underlying objects of interest $\mathbf{X}_k = a^k s_k$ exhibit sparse spectral signatures and sparse spatial morphologies in dictionaries of spectral and spatial waveforms specified *a priori*.

6.2 GMCA for Hyperspectral BSS

6.2.1 Principle

A well known property of the linear mixture model (3) is its *scale and permutation invariance* : without additional prior information, the indexing of the $\mathbf{X}_k$ in the decomposition of data $\mathbf{X}$ is not meaningful and, a^k, s_k can trade a scale factor in full impunity. A consequence is that, unless *a priori* specified otherwise, information on the separate scales of a^k and s_k is lost, and solely a joint scale parameter for a^k, s_k can be estimated. In a Bayesian perspective, this *a priori* knowledge of the multiplicative mixing process and of the loss of information it entails, needs to be translated into a *practical* joint prior probability distribution for $\mathbf{X}_k = a^k s_k$.
The relevant distribution after the multiplicative mixing is the distribution of $\mathbf{X}_k = a^k s_k$, which has the obvious property of being a function of a^k and s_k through their product only. Actually, the variables that matter are γ^k and ν_k which are the sparse coefficient vectors representing respectively a^k in $\mathbf{\Xi}$ and s_k in $\mathbf{\Phi}$:

$$\mathbf{X}_k = \mathbf{\Xi}\alpha_k\mathbf{\Phi} = \mathbf{\Xi}\gamma^k\nu_k\mathbf{\Phi} = a^k s_k \ , \tag{78}$$

where $\alpha_k = \gamma^k\nu_k$ is a rank one matrix of coefficients. For the sake of simplicity $\mathbf{\Phi}$ and $\mathbf{\Xi}$ are two orthonormal bases. Unfortunately, deriving the distribution of the product of two independent random vectors γ^k and ν_k starting from assumptions on their separate distribution functions is notoriously cumbersome. We propose instead that the following p_π is a good candidate *joint* sparse prior distribution for γ^k and ν_k after the loss of information induced

by multiplication :

$$p_\pi(\gamma^k, \nu_k) = p_\pi(\gamma^k \nu_k, 1) \propto \exp(-\lambda_k \|\gamma^k \nu_k\|_1) \propto \exp(-\lambda_k \sum_{i,j} |\gamma^k[i] \nu_k[j]|) \ , \tag{79}$$

where $\gamma^k[i]$ is the i^{th} entry in γ^k and $\nu_k[j]$ is the j^{th} entry in ν_k. The property $\|\gamma^k \nu_k\|_{1,1} = \|\gamma^k\|_1 \|\nu_k\|_1$ is obvious. Thus, the proposed distribution has the nice property, for subsequent derivations, that the conditional distributions of γ^k given ν_k and of ν_k given γ^k are Laplacian distributions which are commonly and conveniently used to model sparse distributions. This distribution provides us with a convenient and formal expression for our prior knowledge of the sparsity of both a^k and s_k in dictionaries of spectral and spatial waveforms and of the multiplicative mixing process. Inserting this prior distribution in a Bayesian maximum a posteriori estimator leads to the following minimization problem :

$$\min_{\{\gamma^k, \nu_k\}} \left\| \mathbf{X} - \sum_{k=1}^{n} \mathbf{\Xi} \gamma^k \nu_k \mathbf{\Phi} \right\|_{\mathbf{\Sigma_N}}^2 + \sum_{k=1}^{n} \lambda_k \|\gamma^k \nu_k\|_1 \ . \tag{80}$$

Interestingly, this can be expressed slightly differently as follows :

$$\min_{\alpha_k} \left\| \mathbf{X} - \sum_{k=1}^{n} \mathbf{X}_k \right\|_{\mathbf{\Sigma_N}}^2 + \sum_{k=1}^{n} \lambda_k \|\alpha_k\|_1 \tag{81}$$

$$\text{with } \ \mathbf{X}_k = \mathbf{\Xi} \alpha_k \mathbf{\Phi} \ \text{ and } \ \forall k, \text{rank}(\mathbf{X}_k) \leq 1$$

thus uncovering a nice interpretation of our problem as that of approximating the data $\mathbf{X}$ by a sum of rank one matrices $\mathbf{X}_k$ which are sparse in the specified dictionary of rank one matrices. This is the usual ℓ_1 minimization problem as in Equation (38), but with the additional constraint that the $\mathbf{X}_k$ are all rank one at most. The latter constraint is enforced here mechanically through a proper parametric representation of $\mathbf{X}_k = a^k s_k$ or $\alpha_k = \gamma^k \nu_k$.
Let us note that rescaling the parameters $\mathbf{A}$ and $\mathbf{S}$ is not as much a problem now as with GMCA, since it does not affect the objective function (80). Indeed, rescaling the columns of the so-called mixing matrix, $\mathbf{A} \leftarrow \rho \mathbf{A}$ while applying the proper inverse scaling to the lines of the source matrix, $\mathbf{S} \leftarrow \frac{1}{\rho} \mathbf{S}$, leaves both the quadratic measure of fit and the ℓ_1 sparsity measure in equation (80) unaltered. Although renormalizing is still worthwhile numerically, it is no longer dictated by the lack of scale invariance of the objective function and the need to stay away from trivial solutions, as in GMCA. In the next section, we will emphasize on the extension of the GMCA algorithm to the hyperspectral BSS issue.

6.2.2 The algorithmic viewpoint

Let us consider now in detail the case where the multichannel dictionary $\mathbf{\Psi} = \mathbf{\Xi} \otimes \mathbf{\Phi}$ is an orthonormal basis obtained as the tensor product of two orthonormal bases $\mathbf{\Xi}$ and $\mathbf{\Phi}$ of respectively spectral and spatial waveforms. As noted previously, when non-unitary or redundant transforms are used, the above are no longer strictly valid. Nevertheless, simple shrinkage still gives satisfactory results in practice as studied in (86; 70). We also assume that $\mathbf{A}$ is left-invertible and that $\mathbf{S}$ is right invertible. In this case, the minimization problem (80) is best formulated in coefficient space, leading to a slightly different however much faster algorithm since there is only one transformation to be applied and this needs to be done only once. For the sake of clarity, we assume that the noise covariance matrix reduces to the $\mathbf{\Sigma_N} = \sigma_{\mathbf{N}}^2 \mathbf{I}$. With these additional assumptions, problem (80) can be rewritten as follows :

$$\min_{\{\gamma^k, \nu_k\}} \frac{1}{{\sigma_{\mathbf{N}}}^2} \left\| \alpha - \sum_{k=1}^{n} \gamma^k \nu_k \right\|_F^2 + \sum_{k=1}^{n} \lambda_k \|\gamma^k\|_1 \|\nu_k\|_1 \ , \tag{82}$$

where we have written $\alpha = \mathbf{\Xi}^T \mathbf{X} \mathbf{\Phi}^T$ the coefficients of the data matrix $\mathbf{X}$ in the multichannel dictionary $\mathbf{\Psi} = \mathbf{\Xi} \otimes \mathbf{\Phi}$. In other words, we are seeking a decomposition of a matrix α into a sum of sparse rank one matrices $\alpha_k = \gamma^k \nu_k$ by minimizing

$$\min_{\gamma, \nu} \frac{1}{{\sigma_{\mathbf{N}}}^2} \|\alpha - \gamma\nu\|_F^2 + \sum_{k=1}^{n} \lambda_k \|\gamma^k \nu_k\|_1 \ , \tag{83}$$

The minimization problem in (83) has at least one solution by coercivity, and is non-convex. But, for fixed γ (resp. ν), the marginal minimization problem over ν (resp. γ) is convex. As solutions of problem (83) have no explicit formulation, we again propose solving it by means of a block-coordinate relaxation iterative algorithm by alternately minimizing with respect to γ holding ν fixed, and vice versa. Thus, by classical ideas in convex analysis, a necessary condition for (γ, ν) to be a minimizer is that the zero is an element of the subdifferential of the objective at (γ, ν). Using (70, Proposition 3.1), this can be written as the system of coupled proximal forward-backward fixed-point equations :

$$\begin{cases} \nu = \Delta_\eta \left(\nu + \frac{1}{{\sigma_{\mathbf{N}}}^2} \boldsymbol{\beta}_\nu \gamma^T \left(\alpha - \gamma\nu \right) \right) \\ \gamma = \Delta_\zeta \left(\gamma + \frac{1}{{\sigma_{\mathbf{N}}}^2} \left(\alpha - \gamma\nu \right) \nu^T \boldsymbol{\beta}_\gamma \right) \end{cases} , \tag{84}$$

where $\boldsymbol{\beta}_\nu$ and $\boldsymbol{\beta}_\gamma$ are relaxation matrices of appropriate sizes such that the spectral radius of $\mathbf{I} - \frac{1}{{\sigma_{\mathbf{N}}}^2} \boldsymbol{\beta}_\nu \gamma^T \gamma$ and $\mathbf{I} - \frac{1}{{\sigma_{\mathbf{N}}}^2} \nu\nu^T \boldsymbol{\beta}_\gamma$ is bounded above by 1. By assumption on left invertibility of $\mathbf{A}$ and the right invertibility of $\mathbf{S}$, $\frac{1}{{\sigma_{\mathbf{N}}}^2} \gamma^T \gamma$ and $\frac{1}{{\sigma_{\mathbf{N}}}^2} \nu\nu^T$ are symmetric and invertible. Hence, taking $\boldsymbol{\beta}_\nu = {\sigma_{\mathbf{N}}}^2 \left(\gamma^T \gamma \right)^{-1}$ and $\boldsymbol{\beta}_\gamma = {\sigma_{\mathbf{N}}}^2 \left(\nu\nu^T \right)^{-1}$, the above are rewritten as the following update rules

on the coefficient matrices γ and ν :

$$\begin{cases} \nu = \Delta_\eta \left(\left(\gamma^T \gamma \right)^{-1} \gamma^T \alpha \right) \\ \gamma = \Delta_\zeta \left(\alpha \nu^T \left(\nu \nu^T \right)^{-1} \right) \end{cases} . \tag{85}$$

η is a vector of size n, each of its entries $\eta[k] = \frac{\lambda_k {\sigma_{\mathbf{N}}}^2 \|\gamma^k\|_{\ell_1}}{2\|\gamma^k\|_{\ell_2}^2}$, and similarly ζ is a vector of size n the entries of which $\zeta[k] = \frac{\lambda_k {\sigma_{\mathbf{N}}}^2 \|\nu_k\|_{\ell_1}}{2\|\nu_k\|_{\ell_2}^2}$. The multichannel soft-thresholding operator Δ_η acts on each row k with threshold $\eta[k]$ and Δ_ζ acts on each column k of γ with threshold $\zeta[k]$.

Both update rules can be interpreted as a soft-thresholding operator applied onto the result of a weighted least-squares regression in the $\mathbf{\Xi} \otimes \mathbf{\Phi}$ representation. Finally, in the spirit of the fast GMCA algorithm described in section 5.3.2, it is proposed here that a solution to the above set of coupled equations (84) can also be approached efficiently using a symmetric iterative alternating least-squares scheme in conjunction with a shrinkage operator with a progressively decreasing threshold. In the present case, the transformation into $\mathbf{\Xi} \otimes \mathbf{\Phi}$ space is applied only once which has a major impact on computation speed, especially when dealing with large hyperspectral datasets. The two stage iterative process leads to the following fast hypGMCA algorithm:

1. Set the number of iterations $I_{\max}$ and initial thresholds $\lambda_k^{(0)}$
2. Transform the data into $\mathbf{X}$ into α
3. While $\lambda_k^{(h)}$ are higher than a given lower bound $\lambda_{\min}$,
   - Update ν assuming γ is fixed:
     $\nu^{(h+1)} = \Delta_{\lambda^{(h)}} \left((\gamma^T \gamma)^{-1} \gamma^T \alpha \right)$
   - Update γ assuming ν is fixed:
     $\gamma^{(h+1)} = \Delta_{\lambda^{(h)}} \left(\alpha \nu^T (\nu \nu^T)^{-1} \right)$
   - Decrease the thresholds $\lambda_k^{(h)}$.
4. Stop when $\lambda_k^{(h)} < \lambda_{\min}$.
5. Transform back the coefficients to get $\mathbf{X} = \mathbf{\Xi} \gamma \nu \mathbf{\Phi}$.

The *coarse to fine* process is again the core of this fast version of GMCA for hyperspectral data. With the threshold successively decreasing towards zero with each iteration, the current sparse approximation is progressively refined by including finer structures alternatingly in the different morphological components, both spatially and spectrally. Here again, soft thresholding results from the use of an ℓ_1 sparsity measure, which as explained earlier comes as a good approximation to the desired ℓ_0 quasi-norm solution. Towards the end of the iterative process, applying a hard threshold instead leads to better results. The final threshold should vanish in the noiseless case or it may be set to a

multiple of the noise standard deviation in the presence of noise as in common detection or denoising methods.

6.3 Comparison with GMCA

Comparison between GMCA and its extension to the hyperspectral case

According to the linear instantaneous mixture model, the data $\mathbf{X}$ are modeled as the linear combination of n sources. In this toy example, the sources will be drawn randomly for a set of 128×128 images featured in Figure 18. The number of drawn sources is $n = 5$. The spectra are generated from a Laplacian probability density with scale parameter $\mu = 1$ in an orthogonal wavelet domain. The spectra are to be positive; note that the GMCA algorithm is flexible enough to account for this assumption. In the next experiments, as we want to assess the impact of the spectral sparsity constraint, we won't take advantage of this prior information. The number of channels is $m = 128$. White Gaussian noise with covariance matrix $\mathbf{\Sigma_N} = \sigma_{\mathbf{N}}^2 \mathbf{I}$ is added.
In the next experiment, we first compare the original GMCA algorithm to its extension for hyperspectal data. This first test will give emphasis on the enhancements provided by the sparse spectral constraint when the signal to noise ratio (SNR) varies from 0 to 40 dB. Figure 19 features 6 out of 128 noisy channels with SNR = 20dB. The GMCA algorithms are computed in the curvelet domain with 100 iterations. Figure 20 depicts the sources estimated by the original GMCA algorithm (panels on the left) and by the GMCA algorithm with spectral sparsity constraints (panels on the left). Visual impression clearly favors the results provided by GMCA with spectral sparsity constraints. More quantitative results are given in Figure 21 which pictures the evolution of the mixing matrix criterion $\mathcal{C}_{\mathbf{A}}$ when the SNR varies from 0 to 40dB. Clearly, accounting for additional prior information provides better recovery results. Furthermore, as shown in Figure 21, the morphospectral sparsity constraint provides more robustness to noise.

Behavior in higher dimensions

In the previous paragraph, we emphasized on the robustness to noise provided by the morphospectral sparsity constraint. Intuitively, for fixed numbers of samples t and channels m, increasing the number of sources entails estimating an increasing number of parameters thus making the separation task more difficult. Accounting for the spectral sparsity assumption should lead to better results when the number of sources increases.
In this 1D toy-example experiment, the entries of $\mathbf{S}$ have been independently

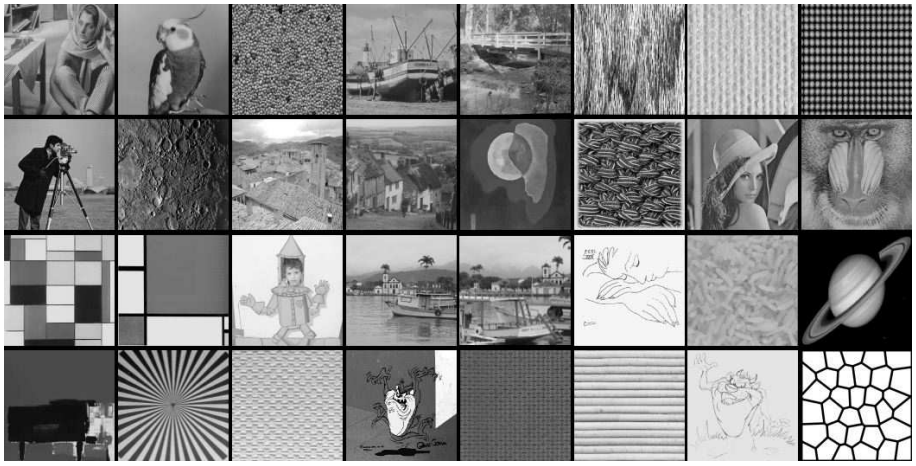

Fig. 18. Image data set used in the experiments.

drawn from a Laplacian probability density with scale parameter $\mu = 1$ ($\mathbf{\Phi}$ is chosen as the Dirac basis). The entries of the mixing matrix are independently drawn from a Laplacian probability density with scale parameter $\mu = 1$ ($\mathbf{\Xi}$ is also chosen as the Dirac basis). The data are not contaminated by noise. The number of samples is $t = 2048$; the number of channels is $m = 128$. Figure 22 depicts the comparisons between GMCA and its extension to the hyperspectral setting. Each point of this figure has been computed as the mean over 100 trials. The panel on the left of Figure 22 features the evolution of the recovery SNR when the number of sources varies from 2 to 64. At low number of sources, the morphospectral sparsity constraint leads to a slight enhancement of the separation results. When the number of sources increases ($n > 15$), the spectral sparsity constraint clearly enhances the recovery results. For instance, when $n = 64$, the GMCA algorithm with the spectral sparsity constraint outperforms the original GMCA up to 12dB. The panel on the right of Figure 22 shows the behavior of the GMCA algorithms with respect to the sparsity-based criterion $\mathcal{C}_{\ell_1}$ introduced in Equation (68). As expected, accounting for the sparsity of the spectra yields sparser results. Furthermore, as the number of sources increases, the deviation between the aforementioned methods becomes wider.

This experiment enlightens the impact of the morphospectral sparsity constraint on the recovery results. As expected, adding further assumptions leads to enhanced performances. In these experiments we illustrated that the morphospectral sparsity constraint yields : i) a better stability with respect to noise contamination, ii) more robustness when the dimensionality (*i.e.* the number of sources) of the problem increases.

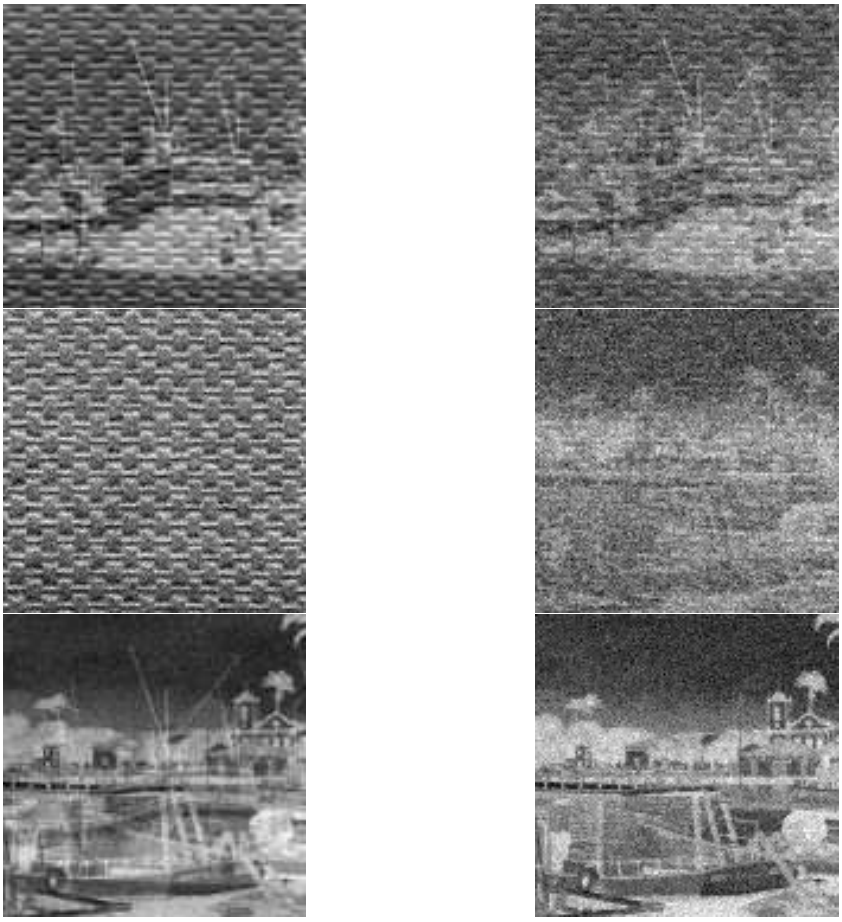

Fig. 19. Six 128×128 mixtures out of the 128 channels. The SNR is equal to 20dB.

7 Applications

7.1 Application to multivalued data restoration

Looking for a sparser representation

In Section 5.5 we enlightened the close links between sparse BSS and best sparsifying/unconditional bases in harmonic analysis. In this context, sparse BSS algorithms are able to provide a basis/representation in which a set of signals (*i.e* the columns of the data matrix in the BSS framework) are **jointly** sparse. In the BSS framework, we proved in Section 5.5 that this nice property leads to the solution of the sparse BSS problem.
Interestingly, in a wide range of multichannel inverse problems, there is a

need for multichannel sparse representations [7] . We illustrated in (81; 98) that looking for an *adaptive* multichannel representation $\mathbf{\Psi} = \mathbf{\Xi} \otimes \mathbf{\Phi}$ in which the data $\mathbf{X}$ are very sparse improves the solution to some classical inverse problems (denoising, inpainting). Let us consider the following general model for inverse problems :

$$\mathbf{Y} = \mathcal{F}(\mathbf{X}) + \mathbf{N} \ . \tag{86}$$

where again $\mathbf{N}$ models noise. The mapping $\mathcal{F}$ is able to represent a variety of degradation operators involved in classical inverse problems (denoising, deconvolution, inpainting to quote a few). We assume that $\mathbf{N}$ is a white Gaussian noise with covariance matrix $\sigma_{\mathbf{N}}^2\mathbf{I}$. We also assume that the data $\mathbf{X}$ are sparse in the multichannel dictionary $\mathbf{\Psi} = \mathbf{\Xi} \otimes \mathbf{\Phi}$. Solving the aforementioned inverse problem is about looking for the solution to the following optimization problem :

$$\min_{\alpha} \lambda \|\alpha\|_1 + \frac{1}{2} \|\mathbf{Y} - \mathcal{F}(\mathbf{\Xi}\alpha\mathbf{\Phi})\|_F^2 \ . \tag{87}$$

Looking for an *adaptive* multichannel representation amounts to adapting the dictionaries $\mathbf{\Xi}$ and $\mathbf{\Phi}$ to the data $\mathbf{X}$. For instance, assuming $\mathbf{\Xi}$ is square invertible, adapting $\mathbf{\Xi}$ is equivalent to seeking a *spectral* representation in which the columns of the data matrix $\mathbf{X}$ are **jointly** sparse. Although no mixture model is available, this task can be performed by applying GMCA on the data $\mathbf{X}$ using $\mathbf{\Phi}$ as the *spatial* dictionary.
Nevertheless, when the data $\mathbf{X}$ are not known up to noise contamination (for instance if $\mathcal{F}$ is different from the identity), estimating $\mathbf{\Xi}$ with GMCA is not possible. In the scope of *adaptive* restoration issues, several approaches can be used :

- Offline scheme : if the data $\mathbf{X}$ are known up to noise contamination (this is the case when $\mathcal{F}$ is the identity mapping), GMCA can be applied on $\mathbf{Y}$ to estimate an appropriate sparser *spectral* representation $\mathbf{\Xi}$. The restoration problem can then be solved assuming $\mathbf{\Xi}$ and $\mathbf{\Phi}$ are fixed. This so-called *offline* scheme is applied for solving a multichannel denoising issue in Section 7.
- Online scheme : if the data $\mathbf{X}$ are degraded by a non-linear mapping $\mathcal{F}$, estimating an adapted *spectral* representation $\mathbf{\Xi}$ cannot be performed using GMCA. We propose to adapt the original GMCA algorithm to solve some inverse problems such as those in Equation (86) while adapting the *spectral* representation $\mathbf{\Xi}$. More precisely, this so-called *online* scheme is applied for solving a multichannel inpainting problem in Section 8.

Adapting the representation to the data has also been introduced in various fields ((99; 100) to quote a few). In (101), Peyré proposed, in the monochannel case, such an adaptive dictionary learning process assuming that the sparse

[7] More precisely, a multichannel dictionary $\mathbf{\Psi} = \mathbf{\Xi} \otimes \mathbf{\Phi}$ as introduced in Section 4 in which the data $\mathbf{X}$ are sparse.

representation lies in a class of tree-based multiscale transforms (e.g. wavelet and cosine packets (71), bandlets (75), etc). Note also that learning patch-based spatial/spectral dictionaries could also be a way of adapting a sparse representation to the data (see (100; 102)).
In the multichannel case, such an adaptive recovery would have to be applied both on the spectral dictionary $\mathbf{\Xi}$ and spatial dictionary $\mathbf{\Phi}$. Note that if the dimension of the data $\mathbf{X}$ are not too high (not exceeding a thousand samples per channel), the GMCA algorithm could be used to adapt the *spatial* dictionary $\mathbf{\Phi}$. In practical situations, the *spatial* dimension t is often much higher than the *spectral* dimension m. In high dimensions, the GMCA algorithm is no longer relevant for adapting sparse representations. The quest for effective *learning* algorithm in high dimensions is still a strenuous and open problem.
In the next section, we assume that the *spatial* dictionary $\mathbf{\Phi}$ is known and fixed. We only look for an adaptive *spectral* dictionary $\mathbf{\Xi}$.

An offline approach - application to color image denoising

In the previous paragraph we emphasized on the crucial importance of signal representations. We claimed that accounting for both *spatial* and *spectral* coherence or structure should enhance multichannel data restoration. In this section, we address the issue of multichannel color image denoising. In this context, the data $\mathbf{X}$ are made of three color layers (Red, Green and Blue). Each color layer is a $\sqrt{t} \times \sqrt{t}$ image. The denoising problem boils down to choosing the perturbation mapping $\mathcal{F}$ in Equation (86) as the identity so that:

$$\mathbf{Y} = \mathbf{X} + \mathbf{N} \tag{88}$$

A first straightforward solution consists in denoising each layer separately. Hopefully, accounting for inter-channel structures or coherence would lead to better results. The top-left picture of Figure 23 portrays a noisy color image with a SNR=15dB. The top-right picture shows the RGB denoised image obtained using a classical wavelet-based denoising method on each color plane [8] using $\mathbf{\Phi}$ as the Undecimated Discrete Wavelet Transform (UDWT).
In Section 5.1, we described a GMCA-based BSS algorithm. We showed that this algorithm is able to seek an adapted *spectral* basis $\mathbf{\Xi}$ in which the processed multichannel data are sparser. We can then apply this algorithm to estimate such a sparse *spectral* representation of different color images. The results are displayed at the bottom of Figure 23. This kind of *learning* step leads to a GMCA-based algorithm that adapts the sparse representation to the data. Such an adaptive process will also be applied to the inpainting problem in Section 8; for further details we refer the reader to (35).
Figure 23 on the right is obtained by applying the GMCA algorithm with the following choices : $\mathbf{\Phi}$ is the UDWT, $\mathbf{\Xi}$ is the adaptive basis obtained with

[8] All color images can be downloaded at *http://perso.orange.fr/jbobin/gmca2.html*.

the GMCA-based BSS algorithm described in 5.1. Visually, denoising in the "adaptive color space" performs better than in the RGB space. Figure 24 zooms on a particular part of the previous images. Visually, the contours are better restored. We also applied this denoising scheme with other non-adaptive *spectral* representations $\mathbf{\Xi}$ (which is equivalent to choosing a different color space representations : YUV, YCC (Luminance and chrominance spaces). For comparative purposes, we also applied the ICA algorithm JADE described in (16) on the original color images to determine yet another adaptive color representation in which to run the same denoising algorithm. A natural question that arises is the following: *is it worth denoising in a different space (YUV, YCC, JADE or GMCA-based) instead of denoising in the original RGB space ?* Figure 25 shows the SNR improvement (in dB) as compared to denoising in the RGB space obtained by each method (YUV, YCC, JADE and GMCA-based). Figure 25 shows that YUV and YCC representations lead to the same results. Note that the YCC color standard is derived from the YUV one. With this particular color image, JADE gives satisfactory results as it can improve denoising up to 1 dB. Finally, as expected, a sparsity-based representation such as the GMCA-based spectral representation provides better results. Here, the use of the sparsest GMCA-based representation enhances denoising up to 2dB. This series of tests confirms the visual impression that we get from Figure 23. According to our claim, accounting for inter-channel coherence improves multichannel data denoising quality.

An online approach - application to color image inpainting

Throughout this paper, we focused on accounting for both *spectral* and *spatial* coherence/structures to better solve multichannel inverse problems such as inpainting or denoising issues. Furthermore, in Section 7, we used the GMCA algorithm to devise a *spectral* basis to better (*i.e.* sparsely) represent the multichannel data. We showed that adapting the representation to the data greatly enhances denoising results. Designing adaptive algorithms is then of crucial importance for restoration issues.

In this section, we consider the particular case of color image inpainting. Again, the data $\mathbf{X}$ consist of 3 observed channels corresponding to each color layer (for instance red, green and blue) which cannot be strictly called spectra. Note that restoring color images in a different color basis (*i.e.* YUV) may sometimes enhance the restoration performance.

We then propose recovering masked color images using the proposed GMCA-inpainting method which seeks to adapt the color space to the data $\mathbf{X}$. In this context, we assume that $\mathbf{\Xi}$ is a 3×3 invertible matrix. In the GMCA

framework, $D' = 1$ and the data $\mathbf{X}$ are the linear combination of D multi-channel morphological components. Adapting the spectral basis $\mathbf{\Xi}$ (*i.e.* the color space) to the data then amounts to estimate an "optimal" matrix $\mathbf{\Xi}$. The GMCA algorithm is then adapted such that at each iteration h the matrix $\mathbf{\Xi}$ is updated by its least-squares estimate:

$$\mathbf{\Xi}^{(h+1)} = \arg\min_{\mathbf{\Xi}} \left\| \mathbf{Y}^{(h)} - \mathbf{\Xi} \sum_{j=1}^{D} \alpha_j^{(h)} \mathbf{\Phi}_j \right\|_F^2 . \tag{89}$$

This problem has a unique minimizer defined as follows:

$$\mathbf{\Xi}^{(h+1)} = \mathbf{Y}^{(h)} \left[\sum_{j=1}^{D} \alpha_j^{(h)} \mathbf{\Phi}_j \right]^{\dagger} , \tag{90}$$

where $\left[\sum_{j=1}^{D} \alpha_j^{(h)} \mathbf{\Phi}_j\right]^{\dagger}$ is the pseudo-inverse of the matrix $\sum_{j=1}^{D} \alpha_j^{(h)} \mathbf{\Phi}_j$.

The GMCA algorithm is then adapted as follows:

1. Set the number of iterations $I_{\max}$ and threshold $\lambda^{(0)}$.
2. While $\lambda^{(h)}$ is higher than a given lower bound $\lambda_{\min}$ (*e.g.* can depend on the noise variance),
   a. Compute $\mathbf{Y}^{(h)} = \mathbf{Y} + \mathcal{M}^c \odot \tilde{\mathbf{X}}^{(h-1)}$.
   b. Initialize to zero each residual morphological components $\{\tilde{\varpi}_j\}^{(h-1)}$.
   For $j = 1, \cdots, D$
   - Compute the residual term $\mathbf{R}_j^{(h)}$ assuming the current estimates of $\varpi_{\{p\}\neq\{j\}}$, $\tilde{\varpi}_{p\neq j}^{(h-1)}$ are fixed:
     $\mathbf{R}_j^{(h)} = \mathbf{Y}^{(h)} - \sum_{p\neq j} \tilde{\varpi}_p^{(h-1)}$.
   - Estimate the current coefficients of $\tilde{\varpi}_j^{(h)}$ by thresholding with threshold $\lambda^{(h)}$:
     $\tilde{\alpha}_j^{(h)} = \Delta_{\lambda^{(h)}} \left(\mathbf{\Xi}^{(h)T} \mathbf{R}_j^{(h)} \mathbf{\Phi}_j^T \right)$.
   - Get the new estimate of ϖ_j by reconstructing from the selected coefficients $\tilde{\alpha}_j^{(h)}$:
     $\tilde{\varpi}_j^{(h)} = \mathbf{\Xi}^{(h)} \tilde{\alpha}_j^{(h)} \mathbf{\Phi}_j$.
   c. Update the hypercube $\tilde{\mathbf{X}}^{(h)} = \sum_{j=1}^{D} \tilde{\varpi}_j^{(h)}$.
   d. Update the *spectral* basis $\mathbf{\Xi}$:
   $\mathbf{\Xi}^{(h+1)} = \mathbf{Y}^{(h)} \left[\sum_{j=1}^{D} \varpi_j^{(h)} \mathbf{\Phi}_j \right]^{\dagger}$.
3. Decrease the threshold $\lambda^{(h)}$ following an appropriate strategy (*e.g.* linear, mMOM).

We remind the reader that the mMOM strategy was described in subsection 4.3.1.

The top-left picture in Figure 26 shows the original *Barbara* color image.

The top-right picture depicts the masked color image where 90% of the color pixels are missing. The bottom-left picture portrays the recovered image using GMCA in the original RGB color space, which amounts to performing a monochannel MCA-based inpainting on each channel; see (103; 104). The last bottom-right picture shows the image recovered with the color space-adaptive GMCA algorithm. The zoom on the recovered images in Figure 27 shows that adapting the color space avoids chromatic aberrations and hence produces a better visual result. This visual impression is quantitatively confirmed by SNR measurements, where the color space-adaptive GMCA improves the SNR by 1dB.

7.2 Application to the Planck data

Introduction to the Planck data set

Investigating Cosmic Microwave Background (CMB) data is of huge scientific importance as it improves our knowledge of the Universe (105). Indeed, most cosmological parameters can be derived from the study of CMB data. In the last decade several experiments (Archeops, Boomerang, Maxima, WMAP - (106)) have already provided large amounts of data and astrophysical information. The forthcoming Planck ESA mission will provide new accurate data requiring effective data analysis tools. More precisely, recovering useful scientific information requires disentangling in the CMB data the contribution of several astrophysical components namely CMB itself, Galactic emissions from dust and synchrotron, Sunyaev-Zel'dovich (SZ) clusters (107) to name a few. In the frequency range used for CMB observations (108), the observed data combines contributions from distinct astrophysical components the recovery of which falls in the frame of component separation.
Following a standard practice in the field of component or source separation, which has physical grounds here, the observed sky is modeled as a linear mixture of statistically independent components. The observation with detector i is then a noisy linear mixture of n independent sources $\{s_j\}_{j=1,\cdots,n}$: $x_i = \sum_{j=1}^{n} a_{ij} s_j + n_i$. The coefficient a_{ij} reflects the emission law of source s_j in the frequency band of the i-th sensor; n_i models instrumental noise.

Applying GMCA to simulations

The GMCA method described above was applied to synthetic data composed of $m = 6$ mixtures of $n = 3$ sources : CMB, galactic dust emission and SZ maps illustrated in Figure 28 and 29. The synthetic data mimic the observations that will be acquired in the six frequency channels of Planck-HFI namely : $100, 143, 217, 353, 545$ and 857 GHz, as shown on Figure 29. White Gaussian

noise $\mathbf{N}$ is added with diagonal covariance matrix $\mathbf{\Sigma_N}$ reflecting the foreseen Planck-HFI noise levels. Experiments were led with 7 *global* noise levels with SNR from 1.7 to 16.7dB such that the experimental noise covariance $\mathbf{\Sigma_N}$ was proportional to the nominal noise covariance. Note that the nominal Planck-HFI global noise level is about 10dB. Each measurement point was computed from 30 experiments involving random noise, randomly chosen sources from a data set of several simulated CMB, galactic dust and SZ 256×256 maps. The astrophysical components and the mixture maps were generated as in (109) according to equation (2) based on model or experimental emission laws, possibly extrapolated, of the individual components. Separation was obtained with GMCA using a single 2D-DWT. Figure 30 depicts the average correlation coefficients over experiments between the estimated source maps and the true source maps. Figure 30 upper left panel shows the correlation coefficient between the true simulated CMB map and the one estimated by JADE (*dotted line with* $\Box$), SMICA (*dashed line with* $\circ$) and GMCA (*solid line*). The CMB map is well estimated by SMICA, which indeed was designed for the blind separation of stationary colored Gaussian processes, but not as well using JADE as one might have expected. GMCA turns out to perform similarly to SMICA. In the second line on the left of Figure 30, galactic dust is well estimated by both GMCA and SMICA. The SMICA estimates seem to have a slightly higher variance than GMCA estimates for higher global noise levels (SNR lower than 5 dB). Finally, the picture in the third line on the left shows that GMCA gives better estimates of the SZ map than SMICA when the noise variance increases. The right panels provide the dispersion (*i.e.* standard deviation) of the correlation coefficients of the sources estimates. It appears that GMCA is a general method yielding simultaneous SZ and CMB estimates comparable to state-of-the-art blind separation techniques which seem mostly dedicated to individual components.
In a noisy context, assessing separation techniques turns out to be more accurate using a mixing matrix criterion, as it is experimentally much more sensitive to separation errors. The bottom right panel of Figure 30 illustrates the behavior of the mixing matrix criterion $\mathcal{C}_{\mathbf{A}}$ with JADE, SMICA and GMCA as the global noise variance varies. GMCA clearly outperforms SMICA and JADE when applied to CMB data.

Adding some physical constraint : the versatility of GMCA

In practice, the separation task is only partly blind. Indeed, the CMB emission law is extremely well-known. In this section, we illustrate that GMCA is versatile enough to account for such prior knowledge. In the following experiment, CMB-GMCA has been designed by constraining the column of the mixing matrix $\mathbf{A}$ related to CMB to its true value. This is equivalent to placing a strict prior on the CMB column of $\mathbf{A}$; that is $P(a^{cmb}) = \delta(a^{cmb} - a_0^{cmb})$ where $\delta(.)$ is the Dirac distribution and a_0^{cmb} is the true simulated CMB emission law

in the frequency range of Planck-HFI. Figure 31 shows the correlation coefficients between the true source maps and the source maps estimated using GMCA with and without the CMB prior. As expected, the top left picture of Figure 31 shows that assuming a_0^{cmb} is known improves the estimation of CMB. Interestingly, the galactic dust map (top right of Figure 31) is also better estimated. Furthermore, the CMB-GMCA SZ map estimate is likely to have a lower variance (lower panel of Figure 31). Moreover, it is likely to provide more robustness to the SZ and galactic dust estimates thus enhancing the global separation performances.

Software

A website have been designed that gives an overview of some applications based on morphological diversity : *http://www.morphologicaldiversity.org*.
A Matlab toolbox coined GMCALab is available online at *http://perso.orange.fr/jbobin/*.

8 Conclusion

In this paper, we overview the application of sparsity and morphological diversity in the scope of blind source separation problems. The contribution of this paper is twofold : (i) it gives new insights into how sparsity enhances blind source separation, (ii) it provides a new sparsity-based source separation method coined Generalized Morphological Component Analysis (GMCA) that takes better advantage of sparsity giving good separation results. GMCA is able to improve the separation task via the use of recent sparse overcomplete (redundant) representations. Numerical results confirm that morphological diversity clearly enhances source separation. When the number of sources is unknown, we introduce a GMCA-based heuristic that provide good separation performances. Further will clearly enlighten the behavior of GMCA when the number of sources has to be estimated. This paper also extends the GMCA framework to the particular case of hyperspectral data. Numerical results are given that illustrates the reliability of morphospectral sparsity constraints.
In a wider framework, GMCA is shown to provide an effective basis for solving classical multivariate restoration problems such as color image denoising or inpainting. Further work will focus on extending GMCA to the underdetermined BSS case (when the number of sources is higher than the number of observations). Finally, GMCA also provides promising prospects in other application such as multivalued data restoration. As GMCA provides a general tool for multivariate data analysis, our future work will also emphasize on the use of GMCA-like methods to other multivalued data applications.

References

[1] J. Tropp, Greedy is good : algorithmic results for sparse approximation, IEEE Transactions on Information Theory 50 (10) (2004) 2231–2242.

[2] A. Jourjine, S. Rickard, O. Yilmaz, Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures., ICASSP '00 5 (2000) 2985–2988.

[3] P. G. Georgiev, F. Theis, A. Cichocki, Sparse component analysis and blind source separation of underdetermined mixtures, IEEE Transactions on Neural Networks 16 (4) (2005) 992–996.

[4] J.-F. Cardoso, Blind signal separation: Statistical principles, Proceedings of the IEEE 86 (1998) 2009–2025.

[5] J.-F. Cardoso, The three easy routes to independent component analysis; contrasts and geometry, in: Proc. ICA 2001, San Diego, 2001.

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

[7] G. Darmois, Analyse générale des liaisons stochastiques, Rev. Inst. Internat. Stat. (1953) 2–8.

[8] J.-F. Cardoso, Dependence, correlation and non Gaussianity in independent component analysis, Journal of Machine Learning Research 4 (2003) 1177–1203.

[9] A. Bell, T. Sejnowski, An information maximisation approach to blind separation and blind deconvolution,, Neural Computation 7 (6) (1995) 1129–1159.

[10] J.-P. Nadal, N. Parga, Non-linear neurons in the low-noise limit: a factorial code maximises information transfer, Network 4 (1994) 295–312.

[11] J.-F. Cardoso, Infomax and maximum likelihood for source separation, IEEE Letters on Signal Processing 4 (4) (1997) 112–114.

[12] B. Pearlmutter, L. Parra, Maximum likelihood blind source separation: A context-sensitive generalization of ica, Advances in Neural Information Processing Systems 9.

[13] D.-T. Pham, P. Garrat, C. Jutten, Separation of a mixture of independent sources through a maximum likelihood approach, in: Proc. EUSIPCO, 1992, pp. 771–774.

[14] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley and Sons, New York, 2001.

[15] A. Belouchrani, K. A. Meraim, J.-F. Cardoso, E. Moulines, A blind source separation technique based on second order statistics, IEEE Trans. on Signal Processing 45 (2) (1997) 434–444.

[16] J.-F. Cardoso, High-order contrasts for independent component analysis, Neural Computation 11 (1) (1999) 157–192.

[17] T.-W. Lee, M. Girolami, A. J. Bell, T. J. Sejnowski, A unifying information-theoretic framework for independent component analysis (1998).

[18] S.-I. Amari, Superefficiency in blind source separation, IEEE Trans. on

Signal Processing 47 (4) (April 1999) 936–944.
[19] S. Amari, J.-F. Cardoso, Blind source separation: semiparametric statistical approach, IEEE Tr. on Signal Processing 45 (11).
[20] A. Cichocki, S. Amari, Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications, John Wiley and Sons, New York, 2002.
[21] P. T. Z. Koldovsky, E. Oja, Efficient variant of algorithm fastica for independent component analysis attaining the cramer-rao lower bound, IEEE Transactions on neural networks 17 (2006) 1265–1277.
[22] M. Davies, Identifiability issues in noisy ica, IEEE Signal processing letters 11 (2004) 470–473.
[23] P. T. Z. Koldovsky, Methods of fair comparison of performance of linear ica techniques in presence of additive noise, No. 5, 2006.
[24] H. Barlow, Possible principles underlying the transformation of sensory messages, Sensory Communications - W.Rosenblith, 1961, pp. 217–234.
[25] B. Olshausen, D. Field, Sparse coding with an overcomplete basis set: A strategy employed by v1?, Vision Research. 37 (2006) 3311–3325.
[26] D. Field, Wavelets, vision and the statistics of natural scenes, Phil. Trans. R. Soc. Lond. A 357 (1999) 2527–2542.
[27] E. Simoncelli, B. Olshausen, Natural image statistics and neural representation, Annual Review of Neuroscience 24 (2001) 1193–1216.
[28] M. Zibulevsky, B. Pearlmutter, Blind source separation by sparse decomposition, Neural Computations 13/4.
[29] M. Zibulevski, Blind source separation with relative newton method, Proccedings ICA2003 (2003) 897–902.
[30] M. Ichir, A. Mohammad-Djafari, Hidden markov models for wavelet-based blind source separation., IEEE Transactions on Image Processing 15 (7) (2006) 1887–1899.
[31] Y. Li, S. Amari, A. Cichocki, D. Ho, S. Xie, Underdetermined blind source separation based on sparse representation, IEEE Transactions on signal processing 54 (2006) 423–437.
[32] A. Bronstein, M. Bronstein, M. Zibulevsky, Y. Zeevi, Sparse ica for blind separation of transmitted and reflected images, Intl. Journal of Imaging Science and Technology (IJIST) 15/1 (2005) 84–91.
[33] E. Vincent, Complex nonconvex ℓ_p norm minimization for underdetermined source separation, in: M. E. Davies, C. J. James, S. A. Abdallah, M. D. Plumbley (Eds.), Independent Component Analysis and Signal Separation, Vol. 4666 of LNCS, Springer, 2007, pp. 430–437.
[34] J. Bobin, Y. Moudden, J.-L. Starck, M. Elad, Morphological diversity and source separation, IEEE Signal Processing Letters 13 (7) (2006) 409–412.
[35] J. Bobin, J.-L. Starck, M. J. Fadili, Y. Moudden, Sparsity and morphological diversity in blind source separation, IEEE Transactions on Image Processing 16 (11) (2007) 2662 – 2674.
URL `http://perso.orange.fr/jbobin/pubs2.html`

[36] J.-L. Starck, E. Candès, D. L. Donoho, The curvelet transform for image denoising, IEEE Transactions on Image Processing 11 (6) (2002) 670–684.

[37] J.-L. Starck, M. Elad, D. Donoho, Image decomposition via the combination of sparse representation and a variational approach, IEEE Transactions on Image Processing 14 (10) (2005) 1570–1582.

[38] Y. Li, S. Amari, A. Cichocki, C. Guan, Underdetermined blind source separation based on sparse representation, IEEE Transactions on information theory 52 (2006) 3139–3152.

[39] J. Bobin, Y. Moudden, J.-L. Starck, Enhanced source separation by morphological component analysis, in: ICASSP '06, Vol. 5, 2006, pp. 833–836.

[40] M. Vetterli, Wavelets, approximation, and compression, IEEE Signal Processing Magazine 18 (5) (2001) 59–73.

[41] D. L. Donoho, M. Vetterli, R. A. DeVore, I. Daubechies, Data compression and harmonic analysis, IEEE Trans. Inform. Theory 44 (1998) 2435–2476.

[42] D. Donoho, I. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, Journal of the American Statistical Association 90 (1995) 1200–1224.

[43] D. L. Donoho, Unconditional bases are optimal bases for data compression and for statistical estimation, Applied and Computational Harmonic Analysis 1 (1993) 100–115.

[44] J.-L. Starck, E. Candès, D. L. Donoho, The curvelet transform for image denoising, IEEE Transactions on Image Processing 11 (6) (2002) 131–141.

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

[46] D. Donoho, X. Huo, Uncertainty principles and ideal atomic decomposition, IEEE Trans. on Inf. Theory 47 (7) (2001) 2845–2862.

[47] S. Chen, D. L. Donoho, M. Saunders, Atomic decomposition by basis pursuit, SIAM Journal on Scientific Computing 20 (1998) 33–61.

[48] D. L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization, Proc. Nat. Aca. Sci. 100 (2003) 2197–2202.

[49] A. Bruckstein, M. Elad, A generalized uncertainty principle and sparse representation in pairs of $\mathbb{R}^N$ bases, IEEE Transactions on Information Theory 48 (2002) 2558–2567.

[50] R. Gribonval, M. Nielsen, Sparse representations in unions of bases, IEEE Transactions on Information Theory 49 (12) (2003) 3320–3325.

[51] J.-J. Fuchs, On sparse representations in arbitrary redundant bases, IEEE Transactions on Information Theory 50 (6) (2004) 1341–1344.

[52] A. Feuer, A. Nemirovsky, On sparse representation in pairs of bases, IEEE Transactions on Information Theory 49 (6) (2003) 1579–1581.

[53] R. Gribonval, M. Nielsen, Sparse representations in unions of bases, IEEE Transactions on Information Theory 49 (12) (2003) 3320–3325.

[54] A. Bruckstein, D. L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM ReviewTo appear.

[55] S. Cotter, B. Rao, K. Engan, K. Kreutz-Delgado, Sparse solutions to linear inverse problems with multiple measurement vectors, IEEE Transactions on Signal Processing 53 (2005) 2477–2488.

[56] M. Fornasier, H. Rauhut, Recovery algorithms for vector valued data with joint sparsity constraints, Preprint - available at *http://www.dsp.ece.rice.edu/cs/*.

[57] J. Chen, X. Huo, Sparse representations for multiple measurement vectors (MMV) in an over-complete dictionary, in: ICASSP '05, 2005.

[58] R. Gribonval, M. Nielsen, Beyond sparsity : recovering structured representations by l_1-minimization and greedy algorithms. – application to the analysis of sparse underdetermined ICA –, Advances in Computational MathematicsIn press.

[59] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Transactions on Signal Processing 41 (12) (1993) 3397–3415.

[60] J. Tropp, A. Gilbert, Signal recovery from partial information via orthogonal matching pursuit, Preprint - available at *http://www.dsp.ece.rice.edu/cs/*.

[61] R. Gribonval, P. Vandergheynst, On the exponential convergence of matching pursuits in quasi-incoherent dictionaries, IEEE Ttrans. Information Theory 52 (1) (2006) 255–261.

[62] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Annals of Statistics 32 (2) (2004) 407–499.

[63] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Statist. Soc. B. 58 (1) (1996) 267–288.

[64] M. R. Osborne, B. Presnell, B. A. Turlach, A new approach to variable selection in least squares problems, IMA Journal of Numerical Analysis 20 (3) (2000) 389–403.

[65] D. M. Malioutov, M. Cetin, A. S. Willsky, Homotopy continuation for sparse signal representation, in: ICASSP '05, Vol. 5, 2005, pp. 733–736.

[66] M. Plumbley, Recovery of sparse representations by polytope faces pursuit, in: ICA06, 2006, pp. 206–213.

[67] D. L. Donoho, Y.Tsaig, Fast solution of ell-1-norm minimization problems when the solution may be sparse., in: Preprint available at http://www.dsp.ece.rice.edu/cs/, 2006.

[68] I. Daubechies, M. Defrise, C. D. Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Comm. Pure Appl. Math 57 (2004) 1413–1541.

[69] M. Figueiredo, R. Nowak, An em algorithm fo wavelet-based image restoration, IEEE Trans. On Image Processing 12 (8) (2003) 906–916.

[70] P. L. Combettes, V. R. Wajs, Signal recovery by proximal forward-

backward splitting, SIAM Journal on Multiscale Modeling and Simulation 4 (4) (2005) 1168–1200.
[71] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1998.
[72] E. Candès, D. Donoho, Ridgelets: the key to high dimensional intermittency?, Philosophical Transactions of the Royal Society of London A 357 (1999) 2495–2509.
[73] E. Candès, D. Donoho, Curvelets, Tech. rep., Statistics, Stanford University (1999).
[74] E. Candès, L. Demanet, D. Donoho, L. Ying, Fast discrete curvelet transforms, SIAM Multiscale Model. Simul 5/3 (2006) 861–899.
[75] E. LePennec, S. Mallat, Sparse geometric image representations with bandelets, IEEE Transactions on Image Processing 14 (4) (2005) 423–438.
[76] M. N. Do, M. Vetterli, The contourlet transform: an efficient directional multiresolution image representation, IEEE Transactions on Image Processing 14 (12) (2005) 2091–2106.
[77] J.-L. Starck, M. Elad, D. L. Donoho, Redundant multiscale transforms and their application for morphological component analysis, Advances in Imaging and Electron Physics 132 (2004) 287–348.
[78] J. Bobin, J.-L. Starck, M. J. Fadili, Y. Moudden, D. L. Donoho, Morphological component analysis: An adaptive thresholding strategy, IEEE Trans. On Image Processing 16 (11) (2007) 2675 – 2681.
[79] D. L. Donoho, Y. Tsaig, I. Drori, J.-L. Starck, Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit, IEEE Transactions On Information TheorySubmitted.
[80] S. Sardy, A. Bruce, P. Tseng, Block coordinate relaxation methods for nonparametric wavelet denoising, Journal of Computational and Graphical Statistics 9 (2) (2000) 361–379.
[81] J. Bobin, Y. Moudden, M. J. Fadili, J.-L. Starck, Morphological diversity and sparsity for multichannel data restoration, Journal of Mathematical Imaging and Vision - in press.
[82] D. Donoho, M. Elad, V. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, IEEE Trans. On Information Theory 52 (2006) 6–18.
[83] J.-J. Fuchs, Recovery conditions of sparse representations in the presence of noise, ICASSP '06 3 (3) (2006) 337–340.
[84] J. Tropp, Just relax: Convex programming methods for subset selection and sparse approximation, IEEE Transactions on Information Theory 52 (3) (2006) 1030–1051.
[85] L. Demanet, L. Ying, Wave atoms and sparsity of oscillatory patterns, ACHAAccepted.
[86] M. Elad, Why simple shrinkage is still relevant for redundant representations?, IEEE Transactions on Information Theory 52 (12) (2006) 5559–5569.
[87] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries,

IEEE Transactions on Signal Processing 41 (12) (1993) 3397–3415.
[88] M. Aharon, M. Elad, A. Bruckstein, k-SVD: An algorithm for designing overcomplete dictionaries for sparse representation, IEEE Transactions on Signal Processing 54 (11) (2006) 4311–4322.
[89] R. Balan, Estimator for number of sources using minimum description length criterion for blind sparse source mixtures, in: M. E. Davies, C. J. James, S. A. Abdallah, M. D. Plumbley (Eds.), Independent Component Analysis and Signal Separation, Vol. 4666 of LNCS, Springer, 2007, pp. 333–340.
[90] H. Akaike, Statistical predictor estimation, Annals Inst. Stat. Math. 22 (1970) 203.
[91] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.
[92] N. Saito, B. Benichou, Sparsity vs. statistical independence in adaptive signal representations: A case study of the spike process, in: G. V. Welland (Ed.), Beyond Wavelets, Studies in Computational Mathematics, Vol. 10/9, 2003, pp. 225–257.
[93] Y. Meyer, Images et vision, Personal Communication.
[94] Wavelab 850 for Matlab7.x, (http://www-stat.stanford.edu/∼wavelab/) (2005).
[95] D. Donoho, High-dimensional data analysis: The curse and blessing of dimensionality, Lecture delivered at the Conference Math Challenges of the 21st Century.
[96] J. Bobin, Gmcalab, http://pagesperso-orange.fr/jbobin/gmcalab.html.
[97] Curvelab 2.1 for Matlab7.x, (http://www.curvelet.org/) (2006).
[98] J. Bobin, M. J. Fadili, Y. Moudden, J.-L. Starck, Morphological diversity and sparsity: New insights into multivariate data analysis, in: Proceedings of the SPIE conference wavelets - SPIE 2007, 2007.
[99] P. Sallee, B. Olshausen, Learning sparse multiscale image representations, Advances in Neural Information Processing Systems 15 (2003) 1327–1334.
[100] J. Mairal, M. Elad, G. Sapiro, Sparse representation for color image restoration, ITIPSubmitted.
[101] G. Peyré, Best basis compressed sensing, in: SSVM, 2007, preprint - available at *http://www.dsp.ece.rice.edu/cs/*.
[102] G. Peyré, Texture synthesis and modification with a patch-valued wavelet transform, in: SSVM 07, 2007.
[103] M. J. Fadili, J.-L. Starck, F. Murtagh, Inpainting and zooming using sparse representations, The Computer Journal - in pressIn press.
[104] M. Elad, J.-L. Starck, D. Donoho, P. Querre, Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA), ACHA 19 (3) (2005) 340–358.
[105] G. Jungman, M. Kamionkowski, A. Kosowsky, D. N. Spergel, Cosmological parameter determination with microwave background maps, Phys. Rev. D 54 (1996) 1332–1344.

[106] C. Bennett, et al., First year wilkinson microwave anisotropy probe (WMAP) observations : preliminary maps and basic results, ApJ. Suppl. 148 (1).

[107] R. Sunyaev, Y. Zel'dovich, The velocity of cluster of galaxies to the microwave background. the possibility of its measurement, Ann. Rev. Astron. Astrophys. 18 (1980) 537.

[108] R. Bouchet, R. Gispert, Foregrounds and cmb experiments: I. semi-analytical estimates of contamination, New Astronomy 4 (443).

[109] Y. Moudden, J.-F. Cardoso, J.-L. Starck, J. Delabrouille, Blind component separation in wavelet space: Application to CMB analysis, Eurasip Journal on Applied Signal Processing 15 (2005) 2437–2454.

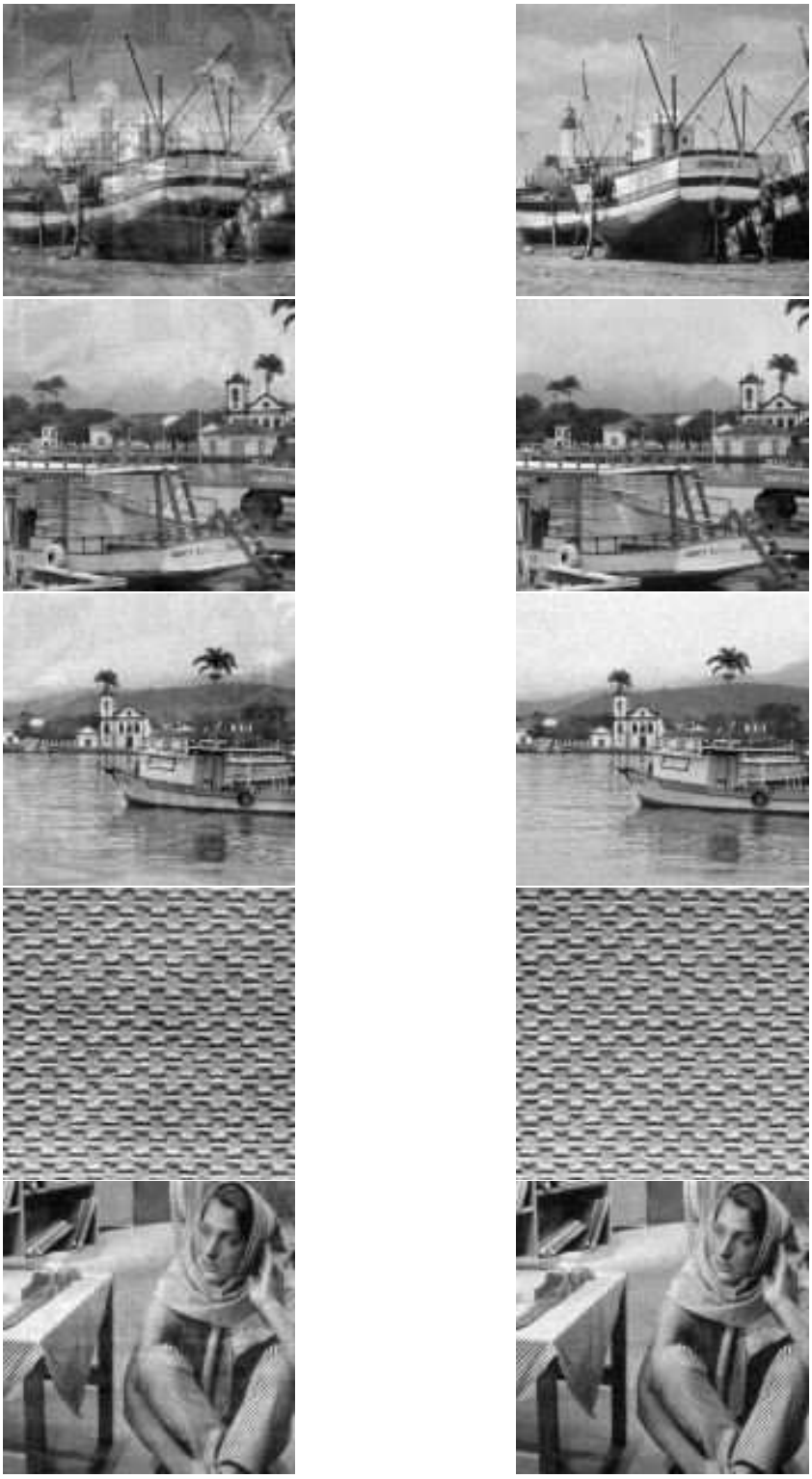

Fig. 20. **Pictures on the left :** Sources estimated with the original GMCA algorithm. **Pictures on the right :** Sources estimated with GMCA with spectral sparsity constraints.

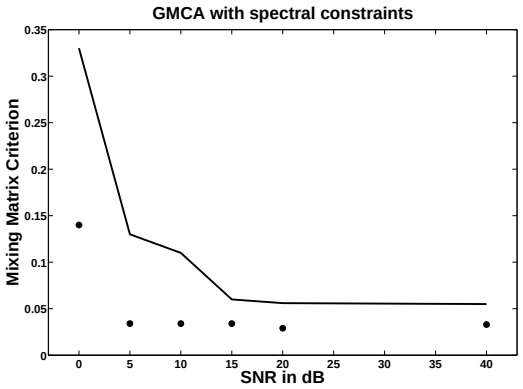

Fig. 21. **Evolution of the mixing matrix criterion $\mathcal{C}_{\mathbf{A}}$ as a function of the SNR in dB.** *Solid line :* recovery results with GMCA. • *:* recovery results with GMCA with spectral sparsity constraint.

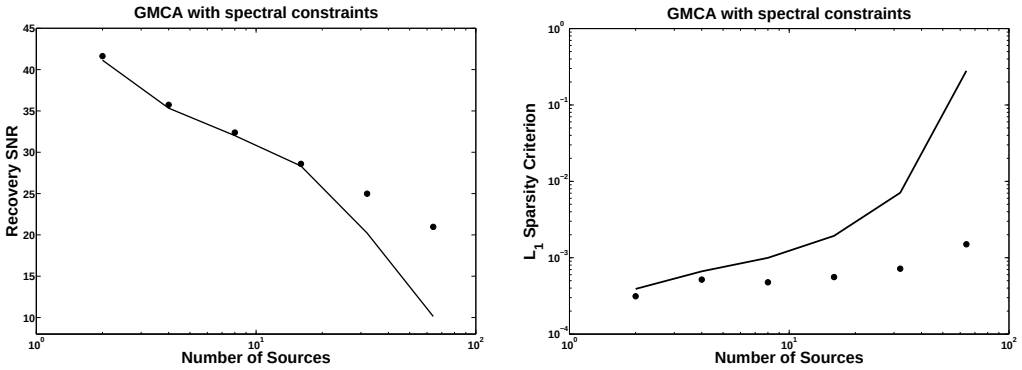

Fig. 22. **Abscissa :** Number of sources. **Ordinate - left :** Recovery SNR. **Right :** sparsity-based criterion $\mathcal{C}_{\ell_1}$. *Solid line :* recovery results with GMCA. • *:* recovery results with GMCA with spectral sparsity constraint.

Fig. 23. **Top-Left :** Original 256×256 image with additive Gaussian noise. The SNR is equal to 15 dB. **Top-Right :** Wavelet-based denoising in the RGB space. **Bottom :** Wavelet-based denoising in the curvelet-GMCA-based spectral representation.

Fig. 24. Zoom the test images. **Top-Left :** Original image with additive Gaussian noise. The SNR is equal to 15 dB. **Top-Right :** Wavelet-based denoising in the RGB space. **Bottom :** Wavelet-based denoising in the curvelet-GMCA space.

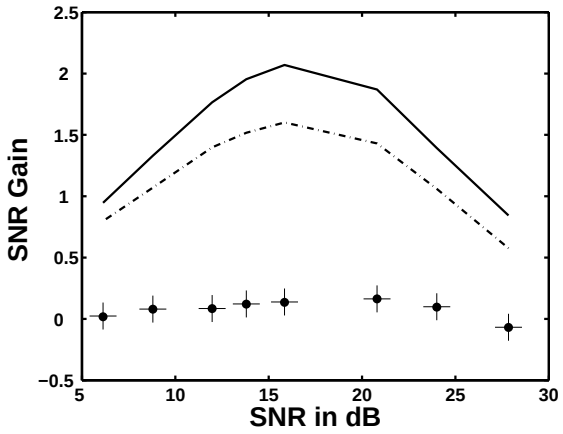

Fig. 25. Denoising color images. **Abscissa :** Mean SNR in dB. **Ordinate :** Gain in terms of SNR in dB compared to a denoising process in the RGB color space. Solid line: GMCA-based, dashed-dotted line: JADE, $'\bullet'$ YUV, $'+'$: YCC.

Fig. 26. Recovering color images. (a) Original Barbara color image. (b) Masked image - 90% of the color pixels are missing. (c) Inpainted image using the original MCA algorithm on each color channel. (d) Inpainted image using the adaptive GMCA algorithm.

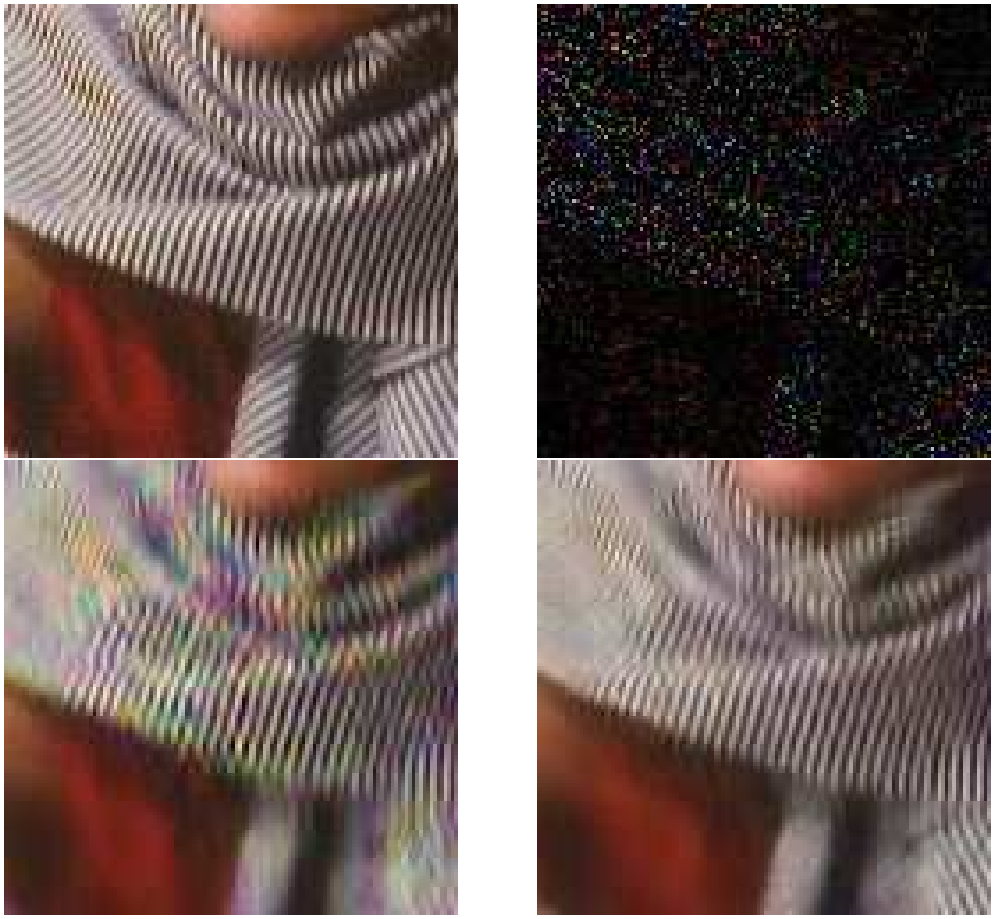

Fig. 27. Zoom on recovered Barbara color image. (a) Original Barbara color image. (b) Masked image - 90% of the color pixels are missing. (c) Inpainted image using the original MCA algorithm on each color channel. (d) Inpainted image using the adaptive GMCA algorithm.

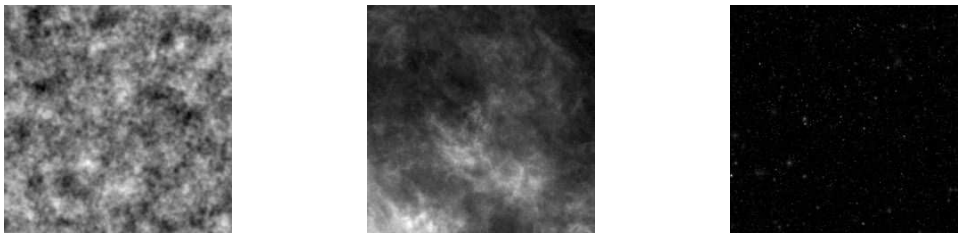

Fig. 28. **The simulated sources - Left:** CMB. **Middle:** galactic dust emission. **Right:** SZ map.

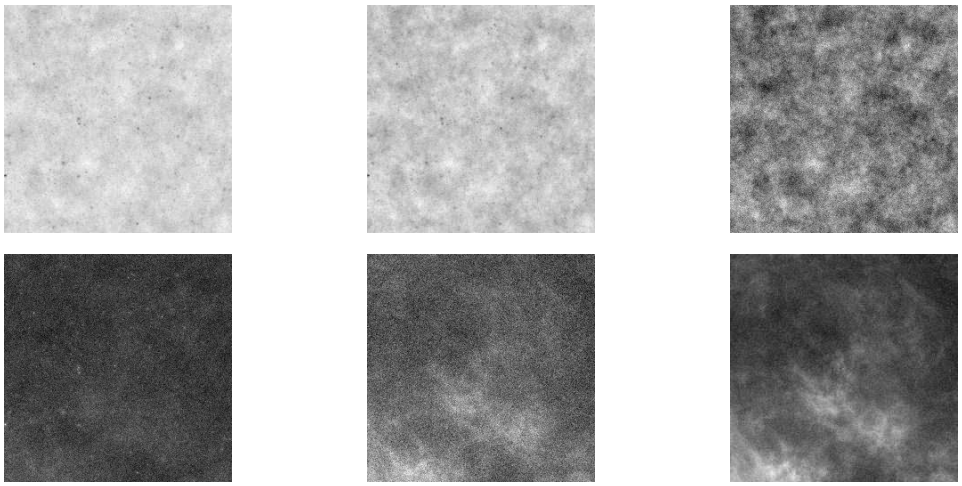

Fig. 29. **The observed CMB data - global SNR = 2.7dB**

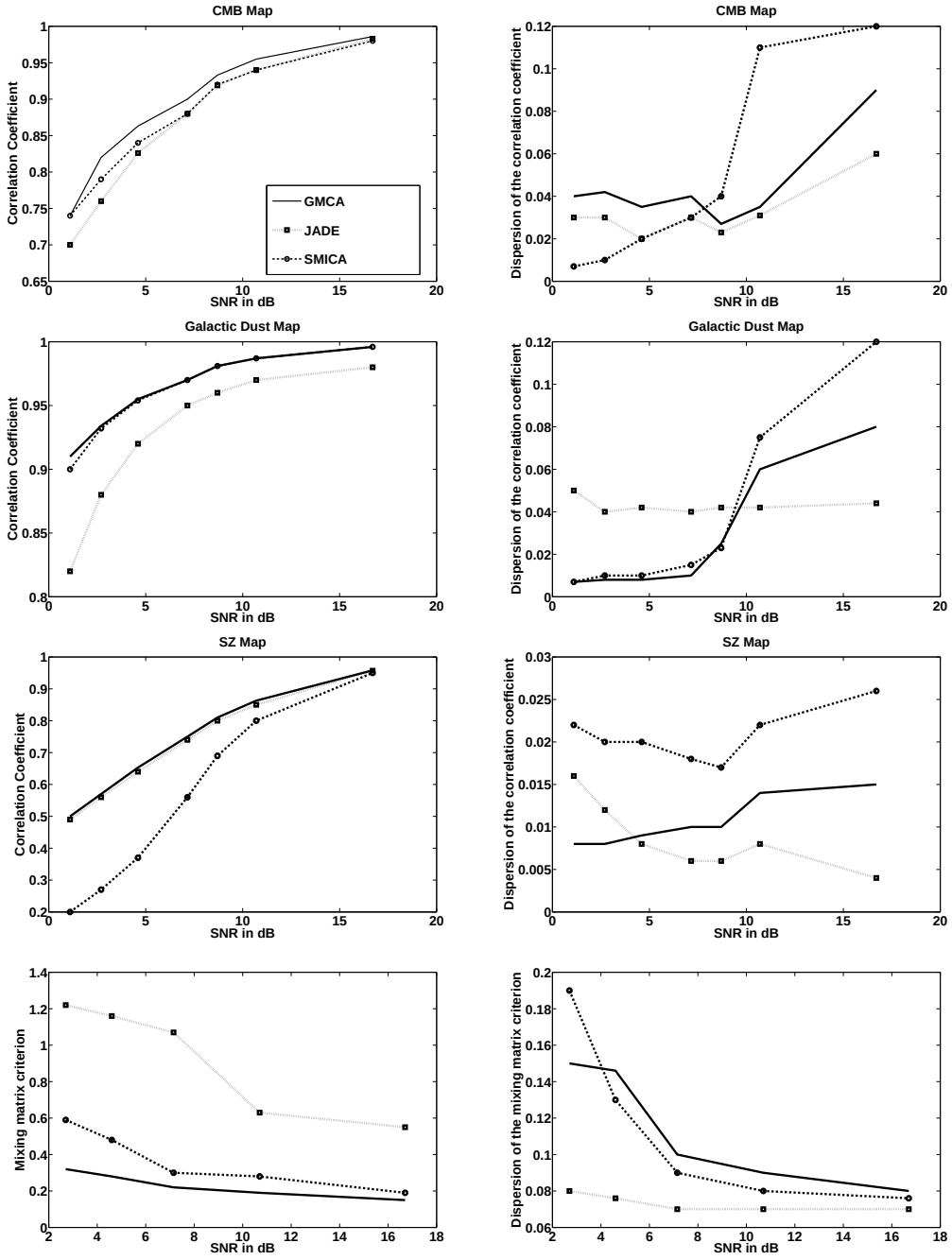

Fig. 30. **Left Column : Correlation coefficients between the estimated source map and the true source map - Left Column : Dispersion of these correlation coefficients : First line :** CMB. **Second line:** galactic dust. **Third line:** SZ map. **Fourth line:** mixing matrix criterion $\mathcal{C}_{\mathbf{A}}$. **Legend :** JADE : *dotted line with* $\square$ - SMICA : *dashed line with* $\circ$ - GMCA : *solid line*. **Abscissa :** SNR in dB.

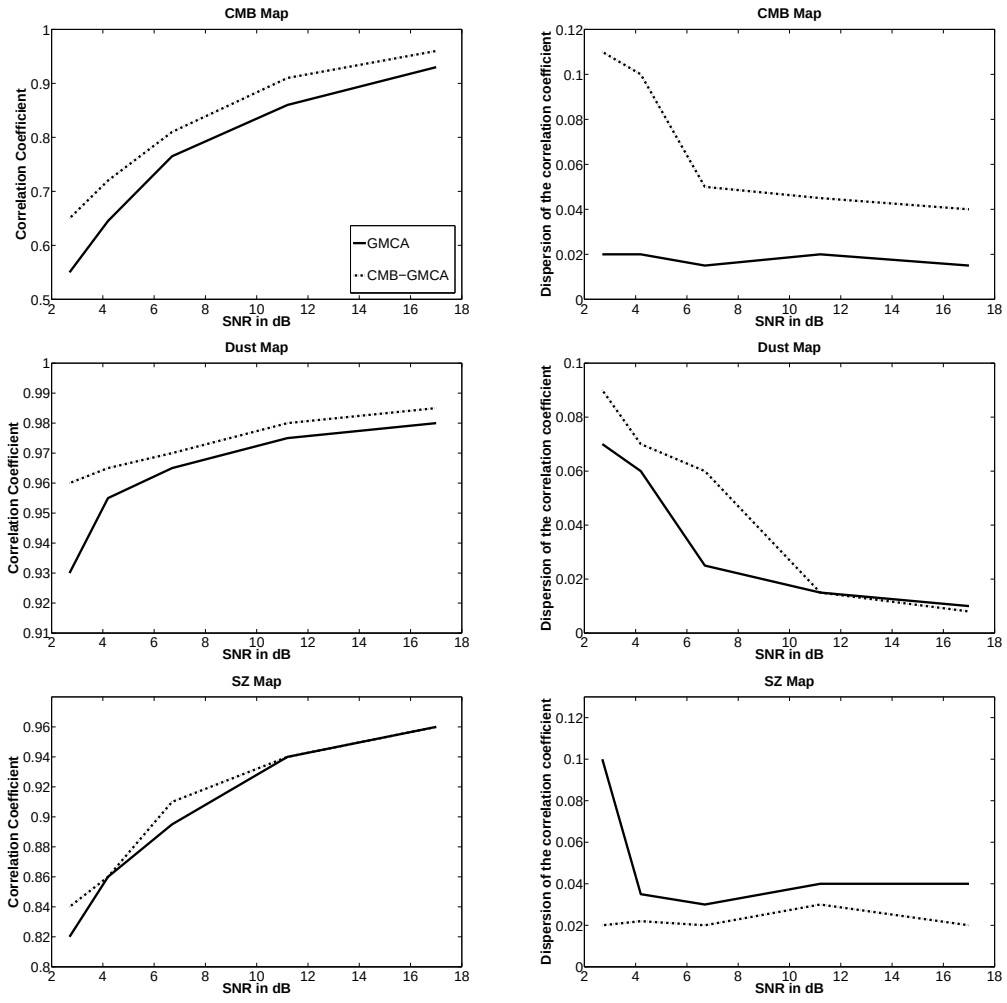

Fig. 31. **Left Column : Mean value of the correlation coefficients between the estimated source map and the true source map - Right Column : Dispersion of these correlation coefficients : First line :** CMB. **Second line:** galactic dust. **Third line:** SZ map. **Legend :** GMCA assuming that the CMB emission law is known : *dotted line* - GMCA : *solid line*. **Abscissa :** SNR in dB.