# Sparsity and Morphological Diversity in Blind Source Separation

J. Bobin*, J.-L. Starck, J. Fadili and Y. Moudden

**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 new and 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 introduces a new 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. We present arguments and a discussion supporting the convergence of the GMCA algorithm. Numerical results in multivariate image and signal processing are given illustrating the good performance of GMCA and its robustness to noise.

EDICS: MRP-WAVL

**Index Terms**

Morphological diversity, sparsity, overcomplete representations, BSS, wavelets, curvelets.

## INTRODUCTION

In the blind source separation (BSS) setting, the instantaneous linear mixture model assumes that we are given $m$ observations $\{x_1, \cdots, x_m\}$ where each $\{x_i\}_{i=1,\cdots,m}$ is a row-vector of size $t$; each measurement

J.Bobin* (E-mail: jerome.bobin@cea.fr), Y.Moudden and J-L.Starck are with the DAPNIA/SEDI-SAP, Service d'Astrophysique, CEA/Saclay, 91191 Gif sur Yvette, France. Phone:+33(0)169083118. Fax:+33(0)169086577.

J-L. Starck is also with Laboratoire APC, 11 place Marcelin Berthelot 75231 Paris Cedex 05, France. E-mail: jstarck@cea.fr.

J. Fadili is with the GREYC CNRS UMR 6072, Image Processing Group, ENSICAEN 14050, Caen Cedex, France. E-mail: Jalal.Fadili@greyc.ensicaen.fr.

is the linear mixture of $n$ source processes :

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

As the measurements are $m$ different mixtures, source separation techniques aim at recovering the original sources $\mathbf{S} = [s_1^T, \cdots, s_n^T]^T$ by taking advantage of some information contained in the way the signals are mixed in the observed data. This mixing model is conveniently 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 the blind approach (where both the mixing matrix $\mathbf{A}$ and the sources $\mathbf{S}$ are unknown), source separation merely boils down to devising quantitative measures of diversity or contrast to differentiate the sources. Most BSS techniques can be separated into two main classes, depending on the way the sources are distinguished:

- Statistical approach - ICA : well-known independent component analysis (ICA) methods assume that the sources $\{s_i\}_{i=1,\cdots,n}$ (modeled as random processes) are statistically independent and non Gaussian. These methods (for example JADE [1], FastICA and its derivatives [2] and [3], Infomax) already provided successful results in a wide range of applications. Moreover, even if the independence assumption is strong, it is in many cases physically plausible. Theoretically, Lee *et al.* [4] emphasize on the equivalence of most of ICA techniques to mutual information minimization processes. Then, in practice, ICA algorithms are about devising adequate contrast functions which are related to approximations of mutual information. In terms of discernibility, statistical independence is a "source of diversity" between the sources.
- Morphological diversity and sparsity : recently, the seminal paper by Zibulevsky *et al.* [5] introduced a novel BSS method that focuses on sparsity to distinguish the sources. They assumed that the sources are sparse in a particular basis $\mathcal{D}$ (for instance orthogonal wavelet basis). The sources $\mathbf{S}$ and the mixing matrix $\mathbf{A}$ are estimated from a Maximum A Posteriori estimator with a sparsity-promoting prior on the coefficients of the sources in $\mathcal{D}$. They showed that sparsity clearly enhances the diversity between the sources. The extremal sparse case assumes that the sources have mutually disjoint

supports (sets of non-zero samples) in the sparse or transformed domain (see [6],[7]). Nonetheless this simple case requires highly sparse signals. Unfortunately this is not the case for large classes of signals and especially in image processing.

A new approach coined Multichannel Morphological Component Analysis (MMCA) is described in [8]. This method is based on **morphological diversity** that is the assumption that the $n$ sources $\{s_i\}_{i=1,\cdots,n}$ we look for are sparse in different representations (i.e. dictionaries). For instance, a piece-wise smooth source $s_1$ (cartoon picture) is well-sparsified in a curvelet tight frame while a warped globally oscillating source $s_2$ (texture) is better represented using a Discrete Cosine Transform (DCT). MMCA takes advantage of this "morphological diversity" to differentiate between the sources with accuracy. Practically, MMCA is an iterative thresholding algorithm which builds on the latest developments in harmonic analysis (ridgelets [9], curvelets [10], [11], [12], etc).

*This paper:* we extend the MMCA method to the much more general case where we consider that each source $s_i$ is a sum of several components ($s_i = \sum_{k=1}^{K} \varphi_k$) each of which is sparse in a given dictionary. For instance, one may consider a mixture of *natural* images in which each is a sum of a piece-wise smooth part (i.e. edges) and a texture component. Using this model, we show that sparsity clearly provides enhancements and gives robustness to noise.

Section I provides an overview of the use of morphological diversity for component separation in single and multichannel images. In section II-A we introduce a new sparse BSS method coined Generalized Morphological Component Analysis (GMCA). Section III shows how to speed up GMCA and the algorithm is described in section IV. In this last section, it is also shown that this new algorithm can be recast as a fixed-point algorithm for which we give heuristic convergence arguments and interpretations. Section V provides numerical results showing the good performances of GMCA in a wide range of applications including blind source separation and multivariate denoising.

## DEFINITIONS AND NOTATIONS

A vector $y$ will be a row vector $y = [y_1, \cdots, y_t]$. Bold symbols represent matrices and $\mathbf{M^T}$ is the transpose of $\mathbf{M}$. The Frobenius norm of $\mathbf{M}$ is $\|\mathbf{Y}\|_2$ defined by $\|\mathbf{Y}\|_2^2 = \text{Trace}\left(\mathbf{Y^T Y}\right)$. The $k$-th entry of $y_p$ is $y_p[k]$, $y_p$ is the $p$-th row and $y^q$ the $q$-th column of $\mathbf{Y}$.

In the proposed iterative algorithms, $\tilde{y}^{(h)}$ will be the estimate of $y$ at iteration $h$. The notation $\|y\|_0$

defines the $\ell_0$ pseudo-norm of $y$ (i.e the number of non-zero elements in $y$) while $\|y\|_1$ is the $\ell_1$ norm of $y$. $\mathcal{D} = [\phi_1^T, \cdots, \phi_T^T]^T$ defines a $T \times t$ dictionary the rows of which are unit $\ell_2$-norm atoms $\{\phi_i\}_i$. The mutual coherence of $\mathcal{D}$ ( see [13] and references therein ) is $\mu_{\mathcal{D}} = \max_{\phi_i \neq \phi_j} |\phi_i \phi_j^T|$. 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 $y$ in $\mathcal{D}$. We thus define $\mathcal{S}_{\ell_0}^{\mathcal{D}}(y)$ (respectively $\mathcal{S}_{\ell_1}^{\mathcal{D}}(y)$) the set of solutions to the minimization problem $\min_c \|c\|_0$ s.t. $y = c\mathcal{D}$ (respectively $\min_c \|c\|_1$ s.t. $y = c\mathcal{D}$). When the $\ell_0$ sparse decomposition of a given signal $y$ has a unique solution, let $\alpha = \Delta_{\mathcal{D}}(y)$ where $y = \alpha\mathcal{D}$ denote this solution. Finally, we define $\lambda_\delta(.)$ to be a thresholding operator with threshold $\delta$ (hard-thresholding or soft-thresholding; this will be specified when needed).

The support $\Lambda(y)$ of row vector $y$ is $\Lambda(y) = \{k; |y[k]| > 0\}$. Note that the notion of support is well-adapted to $\ell_0$-sparse signals as these are synthesized from a few non-zero dictionary elements. Similarly, we define the $\delta$-support of $y$ as $\Lambda_\delta(y) = \{k; |y[k]| > \delta \|y\|_\infty\}$ where $\|y\|_\infty = \max_k |y[k]|$ is the $\ell_\infty$ norm of $y$. In sparse source separation, classical methods assume that the sources have disjoint supports. We define a weaker property for signals $y_p$ and $y_q$ to have $\delta$-disjoint supports if $\Lambda_\delta(y_p) \cap \Lambda_\delta(y_q) = \emptyset$. We further define $\delta^\star = \min\{\delta; \Lambda_\delta(y_p) \cap \Lambda_\delta(y_q) = \emptyset; \quad \forall p \neq q\}$.

Finally, as we deal with source separation, we need a way to assess the separation quality. A simple way to compare BSS methods in a noisy context uses the mixing matrix criterion $\Delta_{\mathbf{A}} = \|\mathbf{I}_n - \mathbf{P}\tilde{\mathbf{A}}^\dagger \mathbf{A}\|_1$, where $\tilde{\mathbf{A}}^\dagger$ is the pseudo-inverse of the estimate of the mixing matrix $\mathbf{A}$, and $\mathbf{P}$ is a matrix that reduces the scale/permutation indeterminacy of the mixing model. Indeed, when $\mathbf{A}$ is perfectly estimated, it is equal to $\tilde{\mathbf{A}}$ up to scaling and permutation. As we use simulations, the true sources and mixing matrix are known and thus $\mathbf{P}$ can be computed easily. The mixing matrix criterion is thus strictly positive unless the mixing matrix is correctly estimated up to scale and permutation.

## I. Morphological Diversity

A signal $y$ is said to be sparse in a waveform dictionary $\mathcal{D}$ if it can be well represented from a few dictionary elements. More precisely, let us define $\alpha$ such that :

$$y = \alpha\mathcal{D} \tag{3}$$

The entries of $\alpha$ are commonly called "coefficients" of $y$ in $\mathcal{D}$. In that setting, $y$ is said to be sparse in $\mathcal{D}$ if most entries of $\alpha$ are nearly zero and only a few have "significant" amplitudes. Particular $\ell_0$-sparse

signals are generated from a few non-zero dictionary elements. Note that this notion of sparsity is strongly dependent on the dictionary $\mathcal{D}$; see e.g. [14], [15] among others. As discussed in [16], 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 harmonic analysis : wavelets, ridgelets [9], curvelets [10], [11], [12], bandlets [17], contourlets [18], 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 large classes of signals. Nevertheless, when $\mathcal{D}$ is overcomplete (*i.e.* $T > t$), the solution of Equation 3 is generally not unique. In that case, the authors of [14] were the first to seek the sparsest $\alpha$, in terms of $\ell_0$-pseudo-norm, such that $y = \alpha\mathcal{D}$. This approach leads to the following minimization problem :

$$\min_{\alpha} \|\alpha\|_0 \text{ s.t. } y = \alpha\mathcal{D} \tag{4}$$

Unfortunately, this is an NP-hard optimization problem which is combinatorial and computationally unfeasible for most applications. The authors of [14] also proposed to convexify the constraint by substituting the convex $\ell_1$ norm to the $\ell_0$ norm leading to the following linear program :

$$\min_{\alpha} \|\alpha\|_1 \text{ s.t. } y = \alpha\mathcal{D} \tag{5}$$

This problem can be solved for instance using interior-point methods. It is known as Basis Pursuit [19] in the signal processing community. Nevertheless, problems (4) and (5) are seldom equivalent. Important research concentrated on finding equivalence conditions between the two problems [15],[20],[21].

In [16] and [22], 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, $y$ is the linear combination of $D$ morphological components:

$$y = \sum_{k=1}^{D} \varphi_k = \sum_{k=1}^{D} \alpha_k \mathbf{\Phi}_k \tag{6}$$

where $\{\mathbf{\Phi}_i\}_{i=1,\cdots,D}$ are orthonormal basis of $\mathbb{R}^t$. Morphological diversity then relies on the sparsity of those morphological components in specific bases. In terms of $\ell_0$ norm, this morphological diversity can be formulated as follows:

$$\forall\{i,j\} \in \{1,\cdots,D\}; \quad j \neq i \Rightarrow \|\varphi_i \mathbf{\Phi}_i^T\|_0 < \|\varphi_i \mathbf{\Phi}_j^T\|_0 \tag{7}$$

In words, MCA then depends 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:

$$\{\varphi_i\} = \text{Arg}\min_{\{\varphi_i\}} \sum_{i=1}^{D} \|\varphi_i \mathbf{\Phi}_i^T\|_1 + \kappa \|y - \sum_{i=1}^{D} \varphi_i\|_2^2 \tag{8}$$

Note that the minimization problem in (8) is closely related to Basis Pursuit Denoising (BPDN - see [19]). In [23], we proposed a particular block-coordinate relaxation, iterative thresholding algorithm (MCA/MOM) to solve (8). 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.

We would like to mention several other methods based on morphological diversity in the specific field of texture/natural part separation in image processing - [24], [25], [26], [27].

In [8], we introduced a multichannel extension of MCA coined MMCA (Multichannel Morphological Component Analysis). In the MMCA setting, we assumed that the sources $\mathbf{S}$ in (2) have strictly different morphologies (*i.e.* each source $s_i$ was assumed to be sparsely represented in one particular orthonormal basis $\mathbf{\Phi}_i$). An iterative thresholding block-coordinate relaxation algorithm was proposed to solve the following minimization problem :

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} \sum_{k=1}^{n} \|s_k \mathbf{\Phi}_k^T\|_1 + \kappa \|\mathbf{X} - \mathbf{A}\mathbf{S}\|_2^2 \tag{9}$$

We then showed in [8] that sparsity and morphological diversity improves the separation task. It confirmed the key role of morphological diversity in source separation to distinguish between the sources.

In the next section, we will introduce a novel way to account for sparsity and morphological diversity in a general Blind Source Separation framework.

## II. Generalized Morphological Component Analysis

### *A. The GMCA framework*

The GMCA framework states that the observed data $\mathbf{X}$ are classically generated as a linear instantaneous mixture of unknown sources $\mathbf{S}$ using an unknown mixing matrix $\mathbf{A}$ as in Equation (2). Note that, we consider here only the overdetermined source separation case where $m \geq n$ and thus $\mathbf{A}$ has full column rank. Future work will be devoted to an extension to the under-determined case $m < n$. An

additive perturbation term $\mathbf{N}$ is added to account for noise or model imperfection. From now, $\mathcal{D}$ is the concatenation of $D$ orthonormal bases $\{\mathbf{\Phi}_i\}_{i=1,\cdots,D}$: $\mathcal{D} = \left[\mathbf{\Phi}_1^T, \cdots, \mathbf{\Phi}_D^T\right]^T$. We assume *a priori* that the sources $\{s_i\}_{i=1,\cdots,n}$ are sparse in the dictionary $\mathcal{D}$. 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 :

$$\forall i \in \{1, \cdots, n\}; \qquad s_i = \sum_{k=1}^{D} \varphi_{ik} = \sum_{k=1}^{D} \alpha_{ik} \mathbf{\Phi}_k \tag{10}$$

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

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} \sum_{i=1}^{n} \sum_{k=1}^{D} \|\varphi_{ik} \mathbf{\Phi}_k^T\|_0 + \kappa \|\mathbf{X} - \mathbf{AS}\|_2^2 \tag{11}$$

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 $\ell_1$ norm to the $\ell_0$ sparsity, which amounts to solving the optimization problem :

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} \sum_{i=1}^{n} \sum_{k=1}^{D} \|\varphi_{ik} \mathbf{\Phi}_k^T\|_1 + \kappa \|\mathbf{X} - \mathbf{AS}\|_2^2 \tag{12}$$

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 $\mathbf{X}_{i,k} = \mathbf{X} - \sum_{\{p,q\} \neq \{i,k\}} a^p \varphi_{pq}$ 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 :

$$\tilde{\varphi}_{ik} = \arg\min_{\varphi_{ik}} \|\varphi_{ik} \mathbf{\Phi}_k^T\|_1 + \kappa \|\mathbf{X}_{i,k} - a^i \varphi_{ik}\|_2^2 \tag{13}$$

or equivalently,

$$\tilde{\alpha}_{ik} = \arg\min_{\alpha_{ik}} \|\alpha_{ik}\|_1 + \kappa \|\mathbf{X}_{i,k} \mathbf{\Phi}_k^T - a^i \alpha_{ik}\|_2^2 \tag{14}$$

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 :

$$0 \in -\frac{1}{\|a^i\|_2^2} a^{i^T} \mathbf{X}_{i,k} \mathbf{\Phi}_k^T + \alpha_{ik} + \frac{1}{2\kappa \|a^i\|_2^2} \partial \|\alpha_{ik}\|_1 \tag{15}$$

where $\partial \|\alpha_{ik}\|_1$ is the subgradient defined as (owing to the separability of the $\ell_1$-norm):

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

Hence, (15) can be rewritten equivalently as two conditions leading to the following closed-form solution:

$$\hat{\alpha}_{jk}[l] = \begin{cases} 0, & \text{if } \left| \left( a^{i^T} \mathbf{X}_{i,k} \mathbf{\Phi}_k^T \right)[l] \right| \leq \frac{1}{2\kappa} \\ \frac{1}{\|a^i\|_2^2} a^{i^T} \mathbf{X}_{i,k} \mathbf{\Phi}_k^T - \frac{1}{2\kappa \|a^i\|_2^2} \text{sign}\left( a^{i^T} \mathbf{X}_{i,k} \mathbf{\Phi}_k^T \right) & \text{otherwise.} \end{cases} \tag{16}$$

This exact solution is known as soft-thresholding. Hence, the closed-form estimate of the morphological component $\varphi_{ik}$ is:

$$\tilde{\varphi}_{ik} = \lambda_\delta \left( \frac{1}{\|a^i\|_2^2} a^{i^T} \mathbf{X}_{i,k} \mathbf{\Phi}_k^T \right) \mathbf{\Phi}_k \text{ with } \delta = \frac{1}{2\kappa \|a^i\|_2^2} \tag{17}$$

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

$$\tilde{a}^i = \frac{1}{\|s_i\|_2^2} \left( \mathbf{X} - \sum_{p \neq i} a^p s_p \right) s_i^T \tag{18}$$

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 [8].

In practice each column of $\mathbf{A}$ is forced to have unit $\ell_2$ 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 variance),

For $i = 1, \cdots, n$

For $k = 1, \cdots, D$

- Compute the residual 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 $\delta^{(h)}$:
  $\tilde{\alpha}_{ik}^{(h)} = \lambda_{\delta^{(h)}} \left( r_{ik}^{(h)} \mathbf{\Phi}_k^T \right)$
- Get the new estimate of $\varphi_{ik}$ 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 thresholds $\delta^{(h)}$.

GMCA is an iterative thresholding algorithm such that at each iteration it first computes *coarse* versions of the morphological component $\{\varphi_{ik}\}_{i=1,\cdots,n;k=1,\cdots,D}$ for a fixed source $s_i$. These raw sources are estimated from their most significant coefficients in $\mathcal{D}$. Hence, the corresponding column $a^i$ 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 $\delta$ decreases towards $\delta_{\min}$. This particular iterative thresholding scheme provides true 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.

### *B. The dictionary $\mathcal{D}$*

As an MCA-like algorithm (for more details, see [8], [23]), the GMCA algorithm involves multiplications by matrices $\mathbf{\Phi}_k^T$ and $\mathbf{\Phi}_k$. Thus, GMCA is worthwhile in terms of computational burden as long as the redundant dictionary $\mathcal{D}$ is a union of bases or tight frames. For such dictionaries, matrices $\mathbf{\Phi}_k^T$ and $\mathbf{\Phi}_k$ are never explicitely constructed, and fast implicit analysis and reconstruction operators are used instead (for instance, wavelet transforms, global or local discrete cosine transform, etc).

### *C. 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 $\phi_{ik}$. 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 III, 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.

### *D. The thresholding strategy*

*Hard or Soft-thresholding ? :* rigorously, we should use a soft-thresholding process. In practice, hard-thresholding leads to better results. Furthermore in [23], we empirically showed that the use of hard-thresholding is likely to provide the $\ell_0$ sparse solution for the single channel sparse decomposition problem. By analogy, we guess that the use of hard-thresholding is likely to solve the multichannel $\ell_0$ norm problem instead of (12).

*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 $\delta$ gets less than $\tau\sigma_{\mathbf{N}}$; $\tau$ 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 an additive noise as a by-product.

### *E. The Bayesian point of view*

We can also consider GMCA from a Bayesian viewpoint. For instance, let's assume that 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's 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 $\mathcal{D}$ (the $\{\alpha_i\}_{i=1,\cdots,n}$) are generated from a Laplacian law:

$$\forall i = 1, \cdots, n; \qquad p(\alpha_i) = \prod_{k=1}^{T} p(\alpha_i[k]) \propto \exp\left(-\mu \|\alpha_i\|_1\right) \tag{19}$$

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

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} \|\mathbf{X} - \mathbf{AS}\|^2_{\mathbf{\Sigma_N}} + 2\mu \sum_{i=1}^{n} \sum_{k=1}^{D} \|\varphi_{ik} \mathbf{\Phi}_k^T\|_1 \tag{20}$$

where $\|.\|_{\mathbf{\Sigma_N}}$ is the Frobenius norm defined such that : $\|\mathbf{X}\|^2_{\mathbf{\Sigma_N}} = \text{Trace}\left(\mathbf{X^T \Sigma_N^{-1} X}\right)$. Note that this minimization task is similar to (11) except that here the metric $\|.\|_{\mathbf{\Sigma_N}}$ accounts for noise. In the case of isotropic and decorrelated noise (*i.e.* $\mathbf{\Sigma_N} = \sigma^2_{\mathbf{N}} \mathbf{I}_m$), problems (12) and (20) are equivalent (with $\kappa = 1/(2\mu\sigma^2_{\mathbf{N}})$).

### *F. Illustrating GMCA*

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 $\mathcal{D}$ 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. Figure 1 illustrates the evolution of $\Delta_{\mathbf{A}}$ as the noise variance decreases. We compare our method to the Relative Newton Algorithm (RNA) [28] that accounts for sparsity and EFICA [3]. The latter is a FastICA variant designed for highly leptokurtotic sources. Both RNA and EFICA were applied after "sparsifying" the data via an orthonormal wavelet transform. Figure 1 shows that GMCA behaves similarly to state-of-the-art sparse BSS techniques.

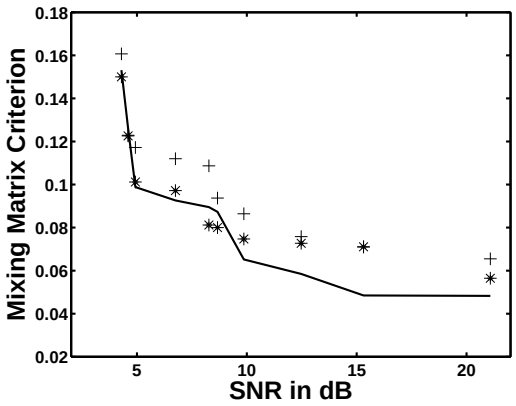


Fig. 1. Evolution of the mixing matrix criterion $\Delta_A$ as the noise variance varies: GMCA (*solid* line), EFICA, *($\star$) :*, RNA *(+)*. **Abscissa :** signal-to-noise ratio in dB. **Ordinate :** mixing matrix criterion value.

## III. Speeding up GMCA

### A. Introduction: the orthonormal case

Let us assume that the dictionary $\mathcal{D}$ is no longer redundant and reduces to an orthonormal basis. The optimization problem (12) then boils down to the following one:

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} \kappa \|\mathbf{\Theta_X} - \mathbf{A}\alpha\|_2^2 + \sum_{i=1}^{n} \|\alpha_i\|_0 \text{ with } \mathbf{S} = \alpha\mathcal{D} \tag{21}$$

where each row of $\mathbf{\Theta_X} = \mathbf{X}\mathcal{D}^T$ stores the decomposition of each observed channel in $\mathcal{D}$. Similarly the $\ell_1$ norm problem (12) reduces to :

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \arg\min_{\mathbf{A},\mathbf{S}} \kappa \|\mathbf{\Theta_X} - \mathbf{A}\alpha\|_2^2 + \sum_{i=1}^{n} \|\alpha_i\|_1 \text{ with } \mathbf{S} = \alpha\mathcal{D} \tag{22}$$

The GMCA algorithm no longer needs transforms at each iteration as only the data $\mathbf{X}$ have to be transformed once in $\mathcal{D}$. 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 (22) to (12) even when the dictionary $\mathcal{D}$ is redundant.

### B. The redundant case

In this section, we assume $\mathcal{D}$ is redundant. We consider that each datum $\{x_i\}_{i=1,\cdots,m}$ has a unique $\ell_0$ sparse decomposition (*i.e.* $\mathcal{S}_{\ell_0}^{\mathcal{D}}(x_i)$ is a singleton for an $i \in \{1, \cdots, m\}$). We also assume that the sources

have unique $\ell_0$ sparse decompositions (*i.e.* $\mathcal{S}_{\ell_0}^{\mathcal{D}}(s_i)$ is a singleton for all $i \in \{1, \cdots, n\}$). We then define $\mathbf{\Theta_X} = [\Delta_{\mathcal{D}}(x_1)^T, \cdots, \Delta_{\mathcal{D}}(x_m)^T]^T$ and $\mathbf{\Theta_S} = [\Delta_{\mathcal{D}}(s_1)^T, \cdots, \Delta_{\mathcal{D}}(s_n)^T]^T$.

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 $\mathcal{D}$. 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 $\mathcal{D}$ : the mixing process increase the $\ell_0$ norm : $\|\Delta_{\mathcal{D}}(x_j)\|_0 > \|\Delta_{\mathcal{D}}(s_i)\|_0$ for all $j \in \{1, \cdots, m\}$ and $i \in \{1, \cdots, n\}$. When $\mathcal{D}$ is made of a single orthogonal basis, this property is exact.
- Sources with $\delta$-disjoint supports in $\mathcal{D}$ : the argument is not so obvious; we guess that the number of significant coefficients in $\mathcal{D}$ is higher for mixture signals than for the original sparse sources with high probability : $\text{Card}\left(\Lambda_\delta(x_j)\right) > \text{Card}\left(\Lambda_\delta(s_i)\right)$ 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 :

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{\Theta}}_\mathbf{S}\} = \arg\min_{\mathbf{A}, \mathbf{\Theta_S}} \kappa \|\mathbf{\Theta_X} - \mathbf{A}\mathbf{\Theta_S}\|_2^2 + \|\mathbf{\Theta_S}\|_0 \tag{23}$$

Obviously, (23) and (11) are not equivalent unless $\mathcal{D}$ is orthonormal. When $\mathcal{D}$ is redundant, no rigorous mathematical proof is easy to derive. Nevertheless, experiments will outline that intuition leads to good results. In (23), note that a key point is still doubtful : sparse redundant decompositions (operator $\Delta_{\mathcal{D}}$) are non-linear and in general no linear model is preserved. Writing $\Delta_{\mathcal{D}}\left(\mathbf{\Theta_X}\right) = \mathbf{A}\Delta_{\mathcal{D}}\left(\mathbf{\Theta_S}\right)$ at the solution is then an invalid statement in general. The next section will focus on this source of fallacy.

### *C. When non-linear processes preserve linearity*

Whatever the sparse decomposition used (e.g. Matching Pursuit [29], Basis Pursuit [19]), 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) :

$$y = \sum_{i=1}^{m} \nu_i y_i \tag{24}$$

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

$$\alpha^\star = \sum_{i=1}^{m} \nu_i \alpha_i \tag{25}$$

As $\alpha^\star$ belongs to $\mathcal{S}_{\ell_0}^{\mathcal{D}}(y)$, a sufficient condition for the $\ell_0$ sparse decomposition to preserve linearity is the uniqueness of the sparse decomposition. Indeed, [14] proved that, in the general case, if

$$||\alpha||_0 < (\mu_{\mathcal{D}}^{-1} + 1)/2 \tag{26}$$

then this is the unique maximally sparse decomposition, and that in this case $\mathcal{S}_{\ell_1}^{\mathcal{D}}(y)$ contains this unique solution as well. Therefore, if all the sources have sparse enough decompositions in $\mathcal{D}$ in the sense of inequality (26), then the sparse decomposition operator $\Delta_{\mathcal{D}}(.)$ preserves linearity.

In [23], the authors showed that when $\mathcal{D}$ is the union of $D$ orthonormal bases, MCA is likely to provide the unique $\ell_0$ pseudo-norm sparse solution to problem (4) when the sources are sparse enough. Furthermore, in [23], experiments illustrate that the Donoho-Huo uniqueness bound is far too pessimistic. Uniqueness should hold, with high probability, beyond the bound (26). Hence, based on this discussion and the results reported in [23], we consider in the next experiments that the operation $\Delta_{\mathcal{D}}(y)$ which stands for the decomposition of $y$ in $\mathcal{D}$ using MCA, preserves linearity.

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

$$x_i = \sum_{j=1}^{n} a_{ij} s_j \tag{27}$$

Owing to the last paragraph, if the sources and the observations have unique $\ell_0$-sparse decompositions in $\mathcal{D}$ then the linear mixing model is preserved, that is:

$$\Delta_{\mathcal{D}}(x_i) = \sum_{j=1}^{n} a_{ij} \Delta_{\mathcal{D}}(s_j) \tag{28}$$

and we can estimate both the mixing matrix and the sources in the sparse domain by solving (23).

## IV. THE FAST GMCA ALGORITHM

According to the last section, a fast GMCA algorithm working in the sparse transformed domain (after decomposing the data in $\mathcal{D}$ using a sparse decomposition algorithm) could be designed to solve (21) (respectively (22)) by an iterative and alternate estimation of $\mathbf{\Theta_S}$ and $\mathbf{A}$. There is an additional important simplification when substituting problem (22) to (12). Indeed, as $m \geq n$, it turns out that (22) is a multichannel *overdetermined* least-squares error fit with $\ell_1$-sparsity penalization. A closely related optimization problem to this augmented lagrangian form is

$$\min_{\mathbf{A},\mathbf{\Theta_S}} \|\mathbf{\Theta_X} - \mathbf{A}\mathbf{\Theta_S}\|_2^2 \quad \text{subject to} \quad \|\mathbf{\Theta_S}\|_1 < q \tag{29}$$

which is a multichannel residual sum of squares with a $\ell_1$-budget constraint. Assuming $\mathbf{A}$ is known, this problem is equivalent to the multichannel fitting regression problem with $\ell_1$-constraint addressed by the homotopy method in [30] or the LARS/Lasso in [31]. While the latter methods are slow stepwise algorithm, we propose the following faster stagewise method:

- Update the coefficients: $\tilde{\mathbf{\Theta}}_\mathbf{S} = \lambda_\delta \left( \tilde{\mathbf{A}}^\dagger \mathbf{\Theta_X} \right)$, $\lambda_\delta$ is a thresholding operator (hard for (21) and soft for (22)) and the threshold $\delta$ decreases with increasing iteration count assuming $\mathbf{A}$ is fixed.
- 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 flavour of the alterning *Sparse coding*/*Dictionary learning* algorithm presented in [32] 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 $\mathbf{\Theta_X}$ :
   $\mathbf{\Theta_X} = [\Delta_\mathcal{D}(x_i)^T]^T$
2. Set the number of iterations $I_{\max}$ and threshold $\{\delta_i^{(0)}\}_{i=1,\cdots,n}$
3. While each $\delta^{(h)}$ is higher than a given lower bound $\delta_{\min}$ (e.g. can depend on the noise variance),
   – Proceed with the following iteration to estimate the coefficients of the sources $\mathbf{\Theta_S}$ at iteration $h$ assuming $\mathbf{A}$ is fixed:
   $\mathbf{\Theta_S}^{(h+1)} = \lambda_{\delta^{(h)}} \left( \mathbf{A}^{\dagger(h)} \mathbf{\Theta_X} \right)$:
   – 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 $\delta^{(h)} = \delta_{\min}$.

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 $\mathcal{D}$. 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 $\delta$ decreases towards a final value $\delta_{\min}$. Similarly to the previous version of the GMCA algorithm (see section II-A), the optimization process provides robustness to noise and helps convergence even in a noisy context. Experiments in Section V illustrate the good performances of our algorithm.

*Complexity analysis:* When the approximation we made is valid, the fast simplified GMCA version requires only the application of MCA on each channel, which is faster than the non-fast version (see Section II-C).

### *A. 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 :

$$\tilde{\mathbf{\Theta}}_{\mathbf{S}} = \lambda_\delta \left( \tilde{\mathbf{A}}^\dagger \mathbf{\Theta_X} \right) \tag{30}$$

where $\tilde{\mathbf{A}}^\dagger$ is the pseudo-inverse of the current estimate $\tilde{\mathbf{A}}$ of the mixing matrix. The next step is a least-squares update of $\mathbf{A}$:

$$\tilde{\mathbf{A}} = \mathbf{\Theta_X} \tilde{\mathbf{\Theta}}_{\mathbf{S}}^{\mathbf{T}} \left( \tilde{\mathbf{\Theta}}_{\mathbf{S}} \tilde{\mathbf{\Theta}}_{\mathbf{S}}^{\mathbf{T}} \right)^{-1} \tag{31}$$

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

$$\tilde{\mathbf{A}} = \tilde{\mathbf{A}} \hat{\mathbf{\Theta}}_{\mathbf{S}} \lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right)^T \left( \lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right) \lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right)^T \right)^{-1} \tag{32}$$

Interestingly, (32) turns out to be a fixed point algorithm. In the next section, we will have a look at its behavior.

*B. Convergence study*

*1) From a deterministic point of view:* A fixed point of the GMCA algorithm is reached when the following condition is verified :

$$\hat{\mathbf{\Theta}}_{\mathbf{S}} \lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right)^T = \lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right) \lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right)^T \tag{33}$$

Note that owing to the non-linear behavior of $\lambda_\delta(.)$ the first term is generally not symmetric as opposed to the second. This condition can thus be viewed as a kind of *symmetrization* condition on the matrix $\hat{\mathbf{\Theta}}_{\mathbf{S}} \lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right)^T$. Let's examine each element of this matrix in the $n = 2$ case without loss of generality. We will only deal with two distinct sources $s_p$ and $s_q$. On the one hand, the diagonal elements are such that:

$$\begin{aligned} \left[ \hat{\mathbf{\Theta}}_{\mathbf{S}} \lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right)^T \right]_{pp} &= \sum_{k=1}^{T} \hat{\alpha}_p[k] \lambda_\delta(\hat{\alpha}_p[k]) \\ &= [\lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right) \lambda_\delta \left( \hat{\mathbf{\Theta}}_{\mathbf{S}} \right)^T]_{pp} \end{aligned} \tag{34}$$

The convergence condition is then always true for the diagonal elements. On the other hand, the off-diagonal elements of (33) are as follows:

$$\sum_{k \in \Lambda_\delta(\hat{\alpha}_q)} \hat{\alpha}_p[k] \hat{\alpha}_q[k] = \sum_{k \in \Lambda_\delta(\hat{\alpha}_p) \cap \Lambda_\delta(\hat{\alpha}_q)} \hat{\alpha}_p[k] \hat{\alpha}_q[k] \text{ and } \sum_{k \in \Lambda_\delta(\hat{\alpha}_p)} \hat{\alpha}_p[k] \hat{\alpha}_q[k] = \sum_{k \in \Lambda_\delta(\hat{\alpha}_p) \cap \Lambda_\delta(\hat{\alpha}_q)} \hat{\alpha}_p[k] \hat{\alpha}_q[k] \tag{35}$$

Let us assume now that the sources have $\delta$-disjoint supports. Define $\delta^\star$ the minimum scalar $\delta$ such that $s_p$ and $s_q$ are $\delta$-disjoint. Similarly, we assume that $\hat{s}_p$ and $\hat{s}_q$ are $\delta$-disjoint and $\delta^\dagger$ is the minimum scalar $\delta$ such that $\hat{s}_p$ and $\hat{s}_q$ are $\delta$-disjoint. Thus for any $\delta > \delta^\dagger$ : $\sum_{k \in \Lambda_\delta(\hat{\alpha}_p) \cap \Lambda_\delta(\hat{\alpha}_q)} \hat{\alpha}_p[k] \hat{\alpha}_q[k] = 0$.

As we noted earlier in Section III-B, when the sources are sufficiently sparse, mixtures are likely to have wider $\delta$-supports than the original sources: $\delta^\star < \delta^\dagger$ unless the sources are well estimated. Thus for any $\delta^\star \leq \delta < \delta^\dagger$ the convergence condition is not true for the off-diagonal terms in (33) as :

$$\sum_{k \in \Lambda_\delta(\hat{\alpha}_q)} \hat{\alpha}_p[k] \hat{\alpha}_q[k] \neq 0 \text{ and } \sum_{k \in \Lambda_\delta(\hat{\alpha}_p)} \hat{\alpha}_p[k] \hat{\alpha}_q[k] \neq 0 \tag{36}$$

Thus the convergence criterion is valid when $\delta^\star = \delta^\dagger$; *i.e.* the sources are correctly recovered up to an "error" $\delta^\star$. When the sources have strictly disjoint supports ($\delta^\star = 0$), the convergence criterion holds true when the estimated sources perfectly match the true sources.

*2) Statistical heuristics:* 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 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 those hypotheses. Figure 2 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 (35):

$$\sum_{k\in\Lambda_\delta(\hat{\alpha}_q)} \mathbb{E}\{\hat{\alpha}_p[k]\hat{\alpha}_q[k]\} = \sum_{k\in\Lambda_\delta(\hat{\alpha}_p)\cap\Lambda_\delta(\hat{\alpha}_q)} \mathbb{E}\{\hat{\alpha}_p[k]\hat{\alpha}_q[k]\} \tag{37}$$

And symmetrically,

$$\sum_{k\in\Lambda_\delta(\hat{\alpha}_p)} \mathbb{E}\{\hat{\alpha}_p[k]\hat{\alpha}_q[k]\} = \sum_{k\in\Lambda_\delta(\hat{\alpha}_p)\cap\Lambda_\delta(\hat{\alpha}_q)} \mathbb{E}\{\hat{\alpha}_p[k]\hat{\alpha}_q[k]\} \tag{38}$$

Intuitively the sources are correctly separated when the branches of the star shaped contour plot (see Figure 2 on the left) of the joint *pdf* of the sources are collinear to the axes.

The question is then: *do conditions (37) and (38) lead to a unique solution* ? *do acceptable solutions belong to the set of fixed points* ? Note that if the sources are perfectly estimated then $\mathbb{E}\{\lambda_\delta\left(\mathbf{\Theta_S}\right)\lambda_\delta\left(\mathbf{\Theta_S}\right)^T\}$

Fig. 2. 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.

is diagonal and $\mathbb{E}\{\mathbf{\Theta_S}\lambda_\delta\left(\mathbf{\Theta_S}\right)\} = \mathbb{E}\{\lambda_\delta\left(\mathbf{\Theta_S}\right)\lambda_\delta\left(\mathbf{\Theta_S}\right)\}$. 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 $\alpha_p$ and $\alpha_q$; hard-thresholding then correlates $\hat{\alpha}_p$ and $\lambda_\delta(\hat{\alpha}_q)$ (respectively $\hat{\alpha_q}$ and $\lambda_\delta(\hat{\alpha}_p)$) unless the joint *pdf* of the estimated sources $\alpha_p$ and $\alpha_q$ has the same symmetries as the thresholding operator (this property has also been outlined in [33]). Figure 3 gives a rather good empirical point of view of the previous remark. On the left, Figure 3 depicts the joint *pdf* of two unmixed sources that have been hard-thresholded. Note that whatever the thresholds we apply, the thresholded sources

Fig. 3. 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.

are still decorrelated as their joint *pdf* verifies the same symmetries as the thresholding operator. On the contrary, on the right of Figure 3, the hard-thresholding process further correlates the two mixtures.

For a fixed $\delta$, several fixed points lead to decorrelated coefficient vectors $\hat{\alpha}_p$ and $\hat{\alpha}_q$. Figure 3 provides a good intuition: for fixed $\delta$ the set of fixed points is divided into two different categories: (i) those which *depend* on the value of $\delta$ (plot on the right) and (ii) those that are valid fixed points for all values of $\delta$ (plot (on the left of Figure 3)). The latter solutions lead to acceptable sources up to scale and permutation. As GMCA involves a decreasing thresholding scheme, the final fixed points are stable if they verify the convergence conditions (37) and (38) for all $\delta$. To conclude, if the GMCA algorithm converges, it should converge to the true sources up to scale and permutation.

### *C. Handling noise*

Sparse decompositions in the presence of noise leads to more complicated results on the support recovery property (see [34] and [35]), and no simple results can be derived for the linearity-preserving property. In practice, we use MCA as a practical sparse signal decomposition. When accounting for noise, MCA is stopped at a given threshold which depends on the noise variance (typically $3\sigma_{\mathbf{N}}$ where $\sigma_{\mathbf{N}}$ is the noise standard deviation). MCA then selects the most significant coefficients of the signal we wish to decompose in $\mathcal{D}$. When the signals are sparse enough in $\mathcal{D}$, such coefficients (with high amplitudes) are less perturbed by noise and thus GMCA provides good results. Indeed, for “very” sparse decompositions with a reasonable signal-to-noise ratio, the influence of noise on the most significant coefficients is rather slight [34]; thus the fixed point property (33) is likely to hold true for most significant coefficients. In that case, “very” sparse decompositions provide robustness to noise. These arguments will be confirmed and supported by the experiments of Section V.

### D. Morphological diversity and statistical independence

In the next section, we give experimental results of comparisons of GMCA to well-known BSS and Independent Component Analysis (ICA) methods. Interestingly, there are close links between ICA and GMCA.

- Theoretically : morphological diversity is, by definition, a deterministic property. As we pointed out earlier, from a probabilistic viewpoint, sources generated independently from a sparse distribution should be morphologically different (i.e. with $\delta$-disjoint support with high probability).
- Algorithmically : we pointed out that GMCA turns to be a fixed point algorithm with convergence condition (33). In [36], the authors present an overview of the ICA fauna in which Equation (33) then turns out to be quite similar to some ICA-like convergence conditions for which a fixed point in B is attained when a matrix $\mathbb{E}\{f\left(\mathbf{BX}\right)\mathbf{BX}^T\}$ is symmetric (in this equation $\mathbf{B}$ is the unmixing matrix and $f(.)$ is the ICA *score* function). In our setting, the operator $\lambda_\delta\left(.\right)$ plays a similar role as the *score* function $f(.)$ in ICA.

In the general case, GMCA will tend to estimate a "mixing" matrix such that the sources are the sparsest in $\mathcal{D}$. We will take advantage of this propensity to look for a multichannel representation (via the estimation of $\mathbf{A}$) in which the estimated components are "very" sparse in $\mathcal{D}$. This point will be illustrated in the next section to denoise color images.

## V. Results

### A. 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 $\mathcal{D}$ (redundant or not), the more "separable" they are. As we noticed earlier, sources with different morphologies are diversely sparse (i.e. they have $\delta$-disjoint supports in $\mathcal{D}$ with a "small" $\delta$). The use of a redundant $\mathcal{D}$ is thus motivated by the grail of sparsity in a wide class of signals for which sparsity means separability.
- **Robustness to noise or model imperfections :** the sparser the sources, the least dramatic the noise. In fact, sparse sources are concentrated on few significant coefficients in the sparse domain for which 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 $1024$ samples (those sources are the *Bump* and *HeaviSine* signals available in the WaveLab toolbox - see [37]). The first column of Figure 4 shows the two synthetic sources. Those sources are randomly mixed so as to provide $m = 2$ observations portrayed by the second column of Figure 4. 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. Gaussian noise with variance $\sigma^2_{\mathbf{N}}$ is added. The third and fourth columns of Figure 4 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 either with a single DWT or with a union of DCT and DWT.

Figure 5 gives the value of the mixing matrix criterion $\Delta_{\mathbf{A}} = \|\mathbf{I}_n - \mathbf{P}\tilde{\mathbf{A}}^{\dagger}\mathbf{A}\|_1$ as the signal-to-noise

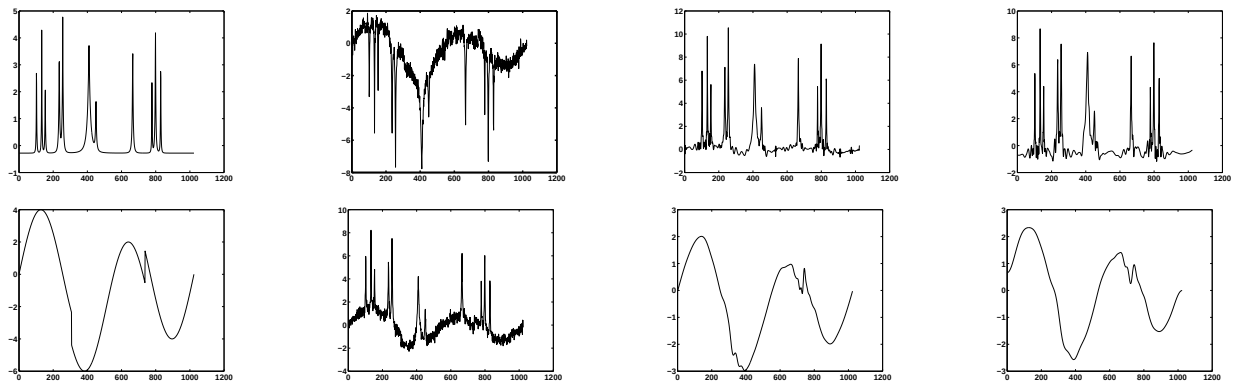

Fig. 4. **The sparser the better - first column:** the original sources. **Second column:** mixtures with additive Gaussian noise (SNR = 19dB). **Third column:** sources estimated with GMCA using a single Discrete Orthogonal Wavelet Transform (DWT). **Fourth column:** Sources estimated with GMCA using a redundant dictionary made of the union of a DCT and a DWT.

ratio (SNR) $10 \log_{10} \left( \|\mathbf{AS}\|_2^2 / \|\mathbf{N}\|_2^2 \right)$ increases. When the mixing matrix is perfectly estimated, $\Delta_{\mathbf{A}} = 0$, otherwise $\Delta_{\mathbf{A}} > 0$. In Figure 5, the *dashed* line corresponds to the behavior of GMCA in a single DWT; the *solid* line depicts the results obtained using GMCA when $\mathcal{D}$ is the union of the DWT and the DCT. On the one hand, GMCA gives satisfactory results as $\Delta_{\mathbf{A}}$ is rather low for each experiment. On the other hand, the values of $\Delta_{\mathbf{A}}$ 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 overcomplete dictionaries.

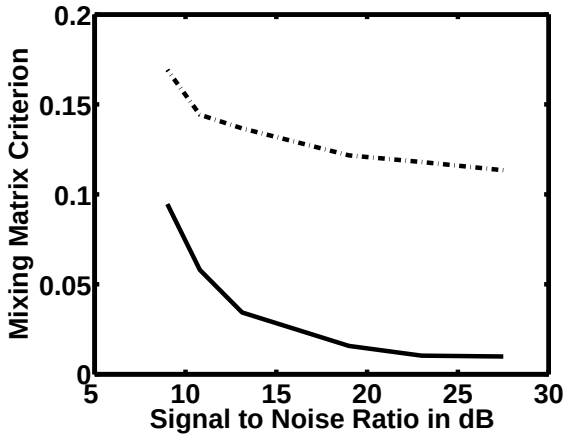


Fig. 5. **The sparser the better :** behavior of the mixing matrix criterion when the noise variance increases for DWT-GMCA (Dashed line) and (DWT+DCT)-GMCA (Solid line).

### *B. 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 [1]).
- Relative Newton Algorithm : the separation technique we already mentioned. This seminal work (see [28]) paved the way for sparsity in Blind Source Separation. In the next experiments, we used the Relative Newton Algorithm (RNA) on the data transformed by a basic orthogonal bidimensional wavelet transform (2D-DWT).
- EFICA : this separation method improves the FastICA algorithm for sources following generalized Gaussian distributions. We also applied EFICA on data transformed by a 2D-DWT where the assumptions on the source distributions is appropriate.

Figure 6 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 the SNR equals 10dB. The noise covariance matrix $\mathbf{\Sigma_N}$ is diagonal.

In section V-A 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 correlation

Fig. 6. **Top :** the $256 \times 256$ source images. **Bottom :** two different mixtures. Gaussian noise is added such that the SNR is equal to 10dB.

coefficient between 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, we also assess the performances of each method by computing the mixing matrix criterion $\Delta_{\mathbf{A}}$. The GMCA algorithm was computed with the union of a Fast Curvelet Transform (available online - see [38], [39]) 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 7 portrays the evolution of the correlation coefficients 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 coefficients closed 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

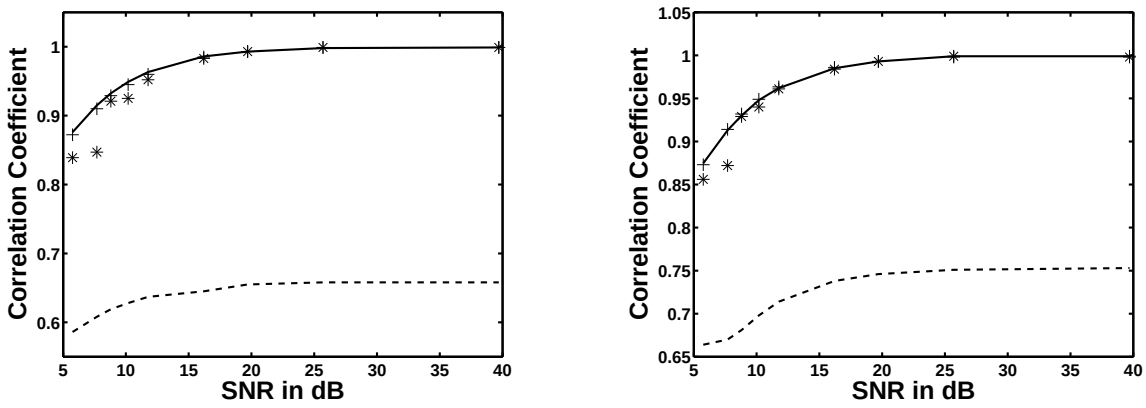


Fig. 7. Evolution of the correlation coefficient between original and estimated sources as the noise variance varies: *solid* line : GMCA, *dashed* line: JADE, *(⋆) :* EFICA, *(+) :* RNA. **Abscissa :** SNR in dB. **Ordinate :** correlation coefficients.

experiments, a mixing matrix-based criterion turns out to be more sensitive to separation errors and then better discriminates between the methods. Figure 8 depicts the behavior of the mixing matrix criterion as the SNR increases. Recall that the correlation coefficients 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 approximately 10 times lower than RNA and 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 outperforms 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 sparse representations than RNA and EFICA.

### *C. Denoising color images*

Up to now we emphasized on sparse blind source separation. Recall that in section IV-B, we showed that the stable solutions of GMCA are the sparsest in the dictionary $\mathcal{D}$. Thus it is tempting to extend GMCA to other multivalued problems such as multi-spectral data restoration.

For instance, it is intuitively appealing to denoise multivalued data (such as color images) in multichannel

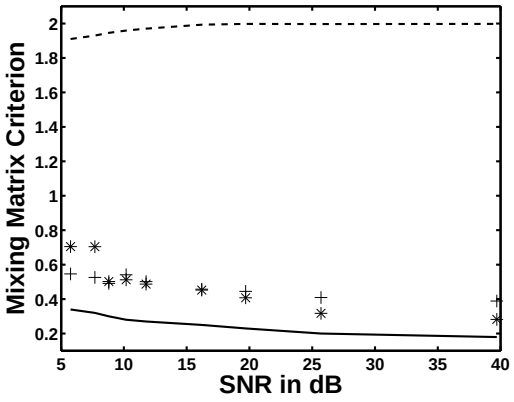


Fig. 8. Evolution of the mixing matrix criterion $\Delta_A$ as the noise variance varies: *solid* line : GMCA, *dashed* line : JADE, *($\star$) :* EFICA, *($+$) :* RNA. **Abscissa :** SNR in dB. **Ordinate :** mixing matrix criterion value.

representations in which the new components are sparse in a given dictionary $\mathcal{D}$. Let's consider multivalued data stored row-wise in the data matrix $\mathbf{X}$. We assume that those multivalued data are perturbed by additive noise. Intuition tells us that it would be worth looking for a new representation $\mathbf{X} = \mathbf{AS}$ such that the new components $\mathbf{S}$ are sparse in the dictionary $\mathcal{D}$. GMCA could be used to achieve this task.

We applied GMCA in the context of color image denoising (SNR = $15$dB). This is illustrated in Figure 9 where the original RGB image[1] are shown on the left. Figure 9 in the middle shows the RGB image obtained using a classical wavelet-based denoising method on each color plane (hard-thresholding in the Undecimated Discrete Wavelet Transform (UDWT)). GMCA is computed in the curvelet domain on the RGB colour channels and the same UDWT-based denoising is applied to the sources $\mathbf{S}$. The denoised data are obtained by coming back to the RBG space via the matrix $\mathbf{A}$. Figure 9 on the right shows the denoised GMCA image using the same wavelet-based denoising method. Visually, denoising in the "GMCA colour space" performs better than in the RGB space. Figure 10 zooms on a particular part of the previous images. Visually, the contours are better restored. Note that GMCA was computed in the curvelet space which is known to sparsely represent piecewise smooth images with $C^2$ contours [10]. We also applied this denoising scheme with other color space representations : YUV, YCC (Luminance and chrominance spaces). We also applied JADE on the original colour images and denoised the components estimated by JADE. The question is then: *would it be worth denoising in a different space (YUV, YCC, JADE or GMCA) instead of denoising in the original RGB space ?* Figure 11 shows the SNR improvement (in dB) as compared to denoising in the RGB space obtained by each method method (YUV, YCC,

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

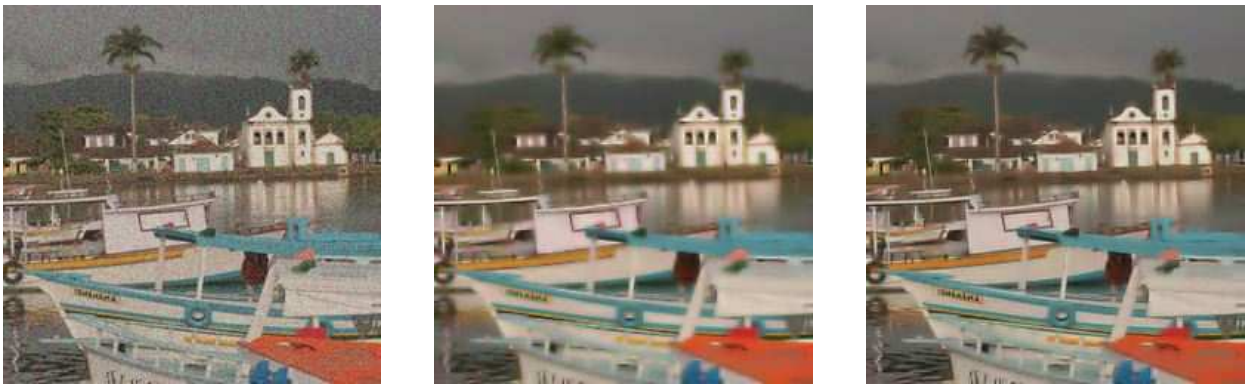

Fig. 9. **Left :** Original $256 \times 256$ image with additive Gaussian noise. The SNR is equal to 15 dB. **Middle :** Wavelet-based denoising in the RGB space. **Right :** Wavelet-based denoising in the curvelet-GMCA space.

JADE and GMCA). Figure 11 shows that YUV and YCC representations lead to the same results. Note that the YCC colour standard is derived from the YUV one. With this particular colour image, JADE gives satisfactory results as it can improve denoising up to 1 dB. Finally, as expected, a sparsity-based representation such as GMCA provides better results. Here, GMCA enhances denoising up to 2dB. This series of tests confirms the visual impression that we get from Figure 9. Note that such "GMCA colour space" is adaptive to the data.

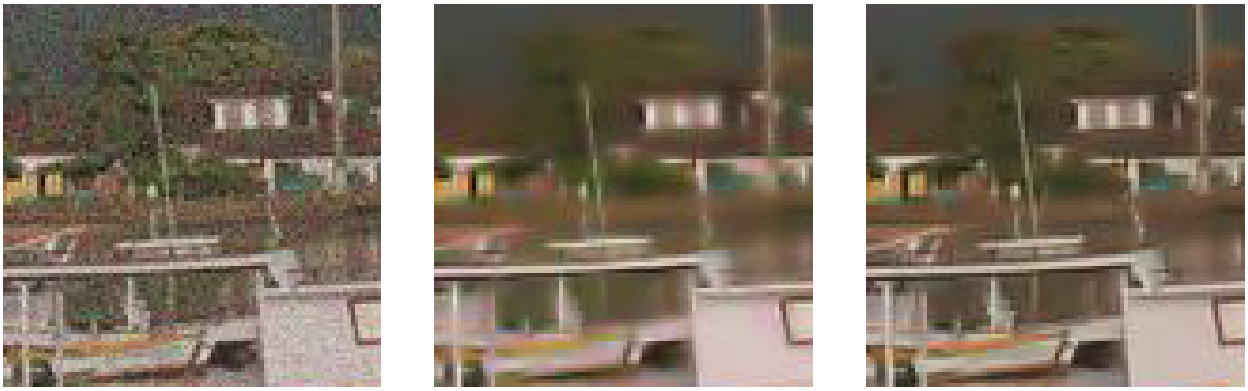

Fig. 10. Zoom the test images. **Left :** Original image with additive Gaussian noise. The SNR is equal to 15 dB. **Middle :** Wavelet-based denoising in the RGB space. **Right :** Wavelet-based denoising in the curvelet-GMCA space.

*a) On the choice of $\mathcal{D}$ and the denoising method :* The denoising method we used is a simple hard-thresholding process in the Undecimated Wavelet (UDWT) representation. Furthermore, $\mathcal{D}$ is a curvelet tight frame (via the fast curvelet transform - [38]). Intuitively, *it would be far better to perform both the estimation of* $\mathbf{A}$ *and denoising in the same sparse representation*. Nonetheless, real facts are much more complicated:

- Estimating the new sparse multichannel representation (through the estimation of $\mathbf{A}$ in $\mathcal{D}$) should

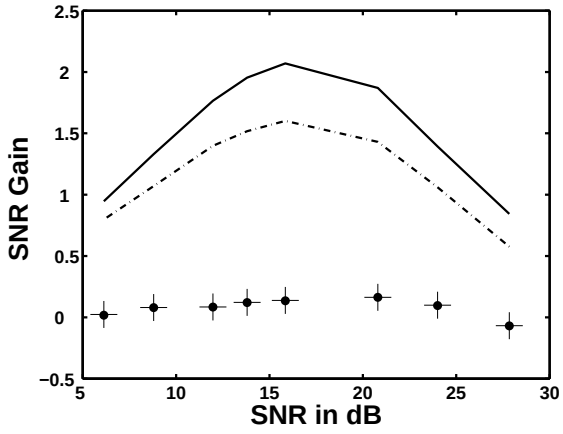


Fig. 11. Denoising color images : how GMCA can improve multivariate data restoration. **Abscissa :** Mean SNR in dB. **Ordinate :** Gain in terms of SNR in dB compared to a denoising process in the RGB color space. Solid line: GMCA, dashed-dotted line: JADE, $'\bullet'$ YUV, $'+'$: YCC.

be performed in the sparsest representation.

- In practice, the "sparsest representation" and the representation for the "best denoising algorithm" are not necessarily identical : (i) for low noise levels, the curvelet representation [38] and the UDWT give similar denoising results. Estimating $\mathbf{A}$ and denoising should give better results in the same curvelet representation, (ii) for higher noise level, UDWT provides a better denoising representation. We then have to balance between (i) *Estimating* $\mathbf{A}$ and (ii) *denoising*; choosing the curvelet representation for (i) and the UDWT for (ii) turns to give good results for a wide range of noise levels.

## SOFTWARE

A Matlab toolbox coined GMCALab will be available online at *http://www.greyc.ensicaen.fr/∼jfadili*.

## VI. CONCLUSION

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. We give conditions under which a simplified GMCA algorithm is designed leading to a fast and effective algorithm. Remarkably, GMCA turns to be equivalent to a fixed point algorithm for which we derive convergence conditions. Our arguments show that GMCA converges to the true sources

up to scale and permutation. Numerical results confirm that morphological diversity clearly enhances source separation. Furthermore GMCA performs well with full benefit of sparsity. Further work will focus on extending GMCA to the under-determined BSS case. Finally, GMCA also provides promising prospects in other application such as multivalued data restoration. Our future work will also emphasize on the use of GMCA-like methods to other multivalued data applications.

## REFERENCES


[1] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE. Special issue on blind identification and estimation*, vol. 9, no. 10, pp. 2009–2025, Oct. 1998.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley, 2001, 481+xxii pages. [Online]. Available: http://www.cis.hut.fi/projects/ica/book/

[3] P. Z.Koldovsky and E.Oja, "Efficient variant of algorithm fastica for independent component analysis attaining the cramer-rao lower bound," *IEEE Transactions on neural networks*, vol. 17, pp. 1265–1277, 2006.

[4] T.-W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski, "A unifying information-theoretic framework for independent component analysis," 1998.

[5] M. Zibulevsky and B.B.Pearlmutter, "Blind source separation by sparse decomposition," *Neural Computations*, vol. 13/4, 2001.

[6] S. A.Jourjine and O.Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures." *ICASSP '00*, vol. 5, pp. 2985–2988, 2000.

[7] Y. Li, S.Amari, A.Cichocki, and C. Guan, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on information theory*, vol. 52, pp. 3139–3152, 2006.

[8] J. Bobin, Y. Moudden, J.-L. Starck, and M. Elad, "Morphological diversity and source separation," *IEEE Signal Processing Letters*, vol. 13, no. 7, pp. 409–412, 2006.

[9] E. Candès and D. Donoho, "Ridgelets: the key to high dimensional intermittency?" *Philosophical Transactions of the Royal Society of London A*, vol. 357, pp. 2495–2509, 1999.

[10] ——, "Curvelets," Statistics, Stanford University, Tech. Rep., 1999.

[11] D. E.Candes, L.Demanet and L.Ying, "Fast discrete curvelet transforms," *SIAM Multiscale Model. Simul*, 2006, to appear.

[12] J.-L. Starck, E. Candès, and D. Donoho, "The curvelet transform for image denoising," *IEEE Transactions on Image Processing*, vol. 11, no. 6, pp. 131–141, 2002.

[13] T. Tropp, "Greed is good : algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[14] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[15] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[16] J.-L. Starck, M. Elad, and D. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1570–1582, 2005.

[17] E. L. Pennec and S. Mallat, "Sparse geometric image representations with bandelets." *IEEE Transactions on Image Processing*, vol. 14, no. 4, pp. 423–438, 2005.

[18] M.N.Do and M.Vetterli, "The contourlet transform: an efficient directional multiresolution image representation." *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2091–2106, 2005.

[19] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[20] A. Bruckstein and M. Elad, "A generalized uncertainty principle and sparse representation in pairs of $\mathbf{r}^n$ bases," *IEEE Transactions on Information Theory*, vol. 48, pp. 2558–2567, 2002.

[21] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

[22] J.-L. Starck, M. Elad, and D. Donoho, "Redundant multiscale transforms and their application for morphological component analysis," *Advances in Imaging and Electron Physics*, vol. 132, pp. 287–348, 2004.

[23] J.Bobin, J-L.Starck, J.Fadili, Y.Moudden, and D.L.Donoho, "Morphological component analysis: new results." *IEEE Transactions on Image Processing - revised - available at http://perso.orange.fr/jbobin/pubs2.html*, 2006.

[24] M. J. Fadili and J.-L. Starck, "Em algorithm for sparse representation - based image inpainting," *IEEE International Conference on Image Processing ICIP'05*, vol. 2, pp. 61–63, 2005, genoa,Italia.

[25] L. A. Vese and S. J. Osher, "Modeling textures with total variation minimization and oscillating patterns in image processing," *UCLA CAM report*, vol. 02-19, 2002.

[26] Y. Meyer, "Oscillating patterns in image processing and in some nonlinear evolution equations." *The Fifteenth Dean Jacquelines B. Lewis Memorial Lectures*, 2001.

[27] J.-F. Aujol, G. Aubert, L. Blanc-Féraud, and A. Chambolle, "Image decomposition into a bounded variation component and an oscillating component." *JMIV*, vol. 22, pp. 71–88, 2005.

[28] M. Zibulevski, "Blind source separation with relative newton method," *Proccedings ICA2003*, pp. 897–902, 2003.

[29] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[30] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, no. 3, pp. 389–403, 2000.

[31] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[32] M.Aharon, M.Elad, and A.Bruckstein, "k-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[33] D.J.Field, "Wavelets, vision and the statistics of natural scenes," *Phil. Trans. R. Soc. Lond. A*, vol. 357, pp. 2527–2542, 1999.

[34] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. On Information Theory*, vol. 52, pp. 6–18, 2006.

[35] J.-J. Fuchs, "Recovery conditions of sparse representations in the presence of noise," *ICASSP '06*, vol. 3, no. 3, pp. 337–340, 2006.

[36] A. Cichocki, “New tools for extraction of source signals and denoising,” in *Proc. SPIE 2005, Bellingham*, vol. 5818, 2005, pp. 11–24.

[37] Wavelab 850 for Matlab7.x, (http://www-stat.stanford.edu/∼wavelab/), 2005.

[38] E.Candes, L.Demanet, D.Donoho, and L.Ying, “Fast discrete curvelet transforms,” *SIAM Multiscale Model. Simul.*, vol. 5/3, pp. 861–899, 2006.

[39] Curvelab 2.01 for Matlab7.x, (http://www.curvelet.org/), 2006.