Blind Separation of Sparse Signals based on Deflation and using the Differential Evolution Algorithm

Henrique E. Oliveira, Leonardo T. Duarte, Yannick Deville, João M. T. Romano

***Abstract*— The aim of this work is the development of a blind source separation method for sparse signals. Our approach is twofold. First, since separation criteria based on the sparsity property often lead to non-convex functions, we address the problem of extracting a single source by performing optimization through a metaheuristic method called differential evolution. Then, a deflation step is set up in order to perform source separation via successive executions of the proposed sparse source extraction algorithm. The resulting method is compared with the gradient descent method by analyzing the existence of local minima in the considered extraction criterion and as well as with respect to the obtained signal-to-interference ratios.**

***Keywords*— Blind Source Separation; Sparsity; Differential Evolution; Deflation.**

I. Introduction

In blind source separation (BSS) [1], [2], [3], the aim is to recover a set of signals (sources) from an ensemble of signals that correspond to mixtures of the sources. This problem finds many applications as, for example, in audio signal separation, telecommunications and hyperspectral imaging [1], [2], [3].

A problem that is closely related to BSS is known as Blind Source Extraction (BSE) [1], [2], [3], in which, instead of recovering all the sources as in BSS, one searches for estimating a single source from the mixtures. A nice aspect of BSE methods is that they can also be applied to perform BSS; this can be done by successive applications of a given BSE method. Note, however, that it becomes necessary to perform a deflation process at the end of each application of the BSE method [4], which consists in removing the contribution of the extracted source from the mixtures.

Among all BSE approaches, the most classical ones are based on the optimization of higher-order statistical cost functions and are strongly connected with independent component analysis (ICA) [5]. More recently, alternative approaches have exploited prior information other than the assumption that the sources are given by independent random variables. For instance, a more recent approach for BSE considers the concept of sparse component analysis (SCA) [6], [7], [8], which is built upon the assumption that the sources are sparse signals.

Henrique E. Oliveira and João M. T. Romano, DSPCom Lab, University of Campinas (UNICAMP), Brazil, Leonardo T. Duarte, School of Applied Sciences, University of Campinas (UNICAMP), Limeira, Brazil, Yannick Deville, Université de Toulouse, UPS-CNRS-OMP, IRAP (Institut de Recherche en Astrophysique et Planétologie), 14 avenue Edouard Belin, F-31400 Toulouse, France. E-mails: henrique@decom.fee.unicamp.br, leonardo.duarte@fca.unicamp.br, romano@fee.unicamp.br, yannick.deville@irap.omp.eu.

In SCA-based approaches, BSE is carried out by the optimization of cost functions based on, for instance, the l_0-pseudo-norm [6], [7], [8], which is one of the most adopted sparsity measures. As already discussed in [6], [9], criteria based on the l_0-pseudo-norm can suffer from local minima and, thus, the application of gradient descent methods may lead to sub-optimal convergence.

In view of this panorama, the aim of this work is to develop an algorithm that perform complete BSS of sparse signals. Our approach considers a metaheuristic method known as differential evolution [10] for performing the optimization of a BSE criterion based on the l_0-pseudo-norm. Moreover, BSS is achieved by successive applications of the developed BSE methods, through a deflation approach that is performed via exhaustive search. The paper is organized as follows. In Section II, we provide a more detailed description of how BSE can be addressed via sparsity-based cost functions. Then, in Section III, we introduce the proposed framework. Numerical experiments are carried out in Section IV, in which the proposed algorithm is compared to the gradient descent. Finally, in Section V, the final conclusions of our work are presented.

II. Blind extraction of sparse sources by the minimization of the l_0-pseudo-norm

A. Extraction criterion

Let us consider a linear mixing system in which each mixture is given by a row of the following matrix:

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

where $\mathbf{A}$ denotes the mixing matrix and each row of the matrix S represents a given source. In this work, we consider determined mixing matrices or, in other words, that the number of mixtures is the same of the number of sources, identified by M and the number of samples is K.

In the problem of BSE [6], the aim is to recover a single source via the following extraction system:

$$\mathbf{y} = \mathbf{w}^T\mathbf{X}, \tag{2}$$

where $\mathbf{w}$ is the extraction vector and $\mathbf{y}$ denotes the retrieved source. A crucial point in BSE, which is related to the adjustment of $\mathbf{w}$, is to define an optimization problem in which the objective function must be a contrast function [1], [3]. For instance, contrast functions based on cumulants were extensively applied for dealing with the case in which the sources can be modeled by independent random variables [2].

More recent works proposed BSE frameworks for the case in which the sources are sparse in [6], [7], [8]. In these scenarios, the objective functions can be built by considering, for instance, l_1-norm [8]. Alternatively, it is also possible to extract a sparse source by minimizing the l_0-pseudo-norm of $\mathbf{y}$, which is simply the number of non-zero elements of $\mathbf{y}$. Indeed, the l_0-pseudo-norm is a contrast function for certain conditions on the sources [6], [9], and, as a consequence, when those conditions are satisfied, BSE can be carried out through the following formulation:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \|\mathbf{y}\|_0 \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 = 1, \end{aligned} \tag{3}$$

where $\|\mathbf{y}\|_0$ denotes the l_0-pseudo-norm of the extracted signal $\mathbf{y}$. The constraint on the l_2-norm of $\mathbf{w}$ must be imposed in order to avoid a trivial solution $\mathbf{y} = \mathbf{0}$, obtained when $\mathbf{w} = \mathbf{0}$.

An alternative, and insightful, definition of the l_0-pseudo-norm of $\mathbf{y}$ is given by

$$\|\mathbf{y}\|_0 = K - \sum_{k=1}^{K} I_0(y_k), \tag{4}$$

where y_k are the components of $\mathbf{y}$ and $I_0(y_k)$ is the indicator function given by

$$I_0(y_k) = \begin{cases} 0, & \text{if } y_k \neq 0 \\ 1, & \text{if } y_k = 0 \end{cases}. \tag{5}$$

This representation opens the way to define smoothed versions of the l_0-pseudo-norm [11]. For instance, if a Gaussian kernel is considered instead of the indicator function, then the following smoothed version of the l_0-pseudo-norm is obtained

$$\|\mathbf{y}\|_0^S = K - \sum_{k=1}^{K} \exp\left(-\frac{y_k^2}{2\sigma^2}\right), \tag{6}$$

where σ corresponds to the standard deviation of the Gaussian kernel. When σ tends to 0, Equation (6) tends to the l_0-pseudo-norm. The rationale behind using a smooth approximation of the l_0-pseudo-norm is that, in real scenarios, sparse signals present many samples that are close but not necessarily equal to zero.

Therefore, a more realistic approach, that shall be considered in the present work, to perform BSE of sparse signals is given by

$$\begin{aligned} \min_{\mathbf{w}} \quad & \left[K - \sum_{k=1}^{K} \exp\left(-\frac{y_k^2}{2\sigma^2}\right)\right] \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 = 1, \end{aligned} \tag{7}$$

where, again, the constraint on $\mathbf{w}$ is to avoid trivial solutions. This constraint is guaranteed simply by dividing the vector $\mathbf{w}$ by its l_2-norm.

B. Optimization by the gradient descent method

The problem expressed in Equation (7) can be solved by a gradient descent method, as follows:

$$w_i^{(n)} = w_i^{(n-1)} - \mu \frac{\partial f}{\partial w_i}^{(n-1)}, \tag{8}$$

where $w_i^{(n)}$ denotes the i-th element of the extraction vector $\mathbf{w}$ at the iteration n, $w_i^{(n-1)}$ denotes the i-th element of the extraction vector $\mathbf{w}$ at the iteration $n-1$ and f denotes, in our case, the smooth approximation of the l_0-pseudo-norm expressed in Equation (6).

It can be easily shown that the gradient of (6) is given by

$$\frac{\partial f}{\partial w_i} = \sum_{k=1}^{K} \frac{y_k x_{ik}}{\sigma^2} \exp\left(-\frac{y_k^2}{2\sigma^2}\right), \tag{9}$$

where x_{ik} denotes the k-th sample of the i-th mixture.

Therefore, the learning rule in this case is given by:

$$w_i^{(n)} = w_i^{(n-1)} - \mu \left[\sum_{k=1}^{K} \frac{y_k x_{ik}}{\sigma^2} \exp\left(-\frac{y_k^2}{2\sigma^2}\right)\right]^{(n-1)}. \tag{10}$$

III. PROPOSED BSS FRAMEWORK

A. Optimization by the differential evolution method

Differential evolution is a metaheuristic method developed by Storn and Price [10] that, as the authors showed, converges faster and gives more accurate results than some traditional methods. Besides that, we choose this method for this work because it has few control variables and is robust, that is, the variables are easy to control [10]. Adapted to the minimization of l_0-pseudo-norm, it has the following structure:

1) Initialization of the population
   First, a matrix $\mathbf{W}$ is initialized in the following way:
   $$\mathbf{W} = [\mathbf{w}_p], \tag{11}$$
   where $\mathbf{w}_p$ is a randomly initialized column vector, with dimension M (number of the mixtures) and $1 \leq p \leq NP$ (population size).
2) Mutation of the vectors $\mathbf{w}_p$
   In this step, mutant vectors $\mathbf{v}_p$ are generated by:
   $$\mathbf{v}_p = \mathbf{w}_{r1} + F(\mathbf{w}_{r2} - \mathbf{w}_{r3}), \tag{12}$$
   where r_1, r_2 and r_3 are different randomly selected indices, F is the mutation rate, a real and constant factor chosen in the range $[0, 2]$, that scales the amplitude of the difference $(\mathbf{w}_{r2} - \mathbf{w}_{r3})$ and $1 \leq p \leq NP$.

3) Crossover between the vectors $\mathbf{w}_p$ and $\mathbf{v}_p$
After, to increase the variability of the vectors, a crossover is performed, where each element i of the new vector is given by:

$$u_{ip} = \begin{cases} v_{ip}, & \text{if } r(i) \leq CR \text{ or } i = s(p) \\ w_{ip}, & \text{otherwise} \end{cases}, \quad (13)$$

where $r(i)$ is a random number in the range $[0, 1]$, $s(p)$ is an index randomly chosen between 1 and M, CR is the crossover rate chosen in the range $[0, 1]$, $1 \leq p \leq NP$ and $1 \leq i \leq M$.
4) Evaluation of the cost function
Then, for each vector $\mathbf{u}_p$ generated after the mutation and the crossover, the cost function of the optimization problem is evaluated by the Equation (6).
5) Selection
Finally, each original element of the population $\mathbf{w}_p$ is compared with the associated modified vector $\mathbf{u}_p$, and the one that leads to the smaller value of the cost function is chosen to become the new $\mathbf{w}_p$.

Despite the excellent results obtained by the differential evolution method, it has limitations when ones increases the numbers of parameters, in more complex applications [10].

B. A BSS approach based on deflation

As discussed before, we can perform BSS by executing several times a BSE algorithm and by removing, at the end of each execution, the contribution of the extracted source from the mixtures [4]. For example, having extracted $\mathbf{y}_1$, which represents the first extracted signal, the deflation procedure is done as follows:

$$\mathbf{z}_i = \mathbf{x}_i - \alpha_i \mathbf{y}_1, \quad (14)$$

where $\mathbf{x}_i$ denotes the i-th original mixture and $\mathbf{z}_i$ corresponds to the new mixture i, ideally free of contributions from the source that was estimated by $\mathbf{y}_1$. This procedure is repeated for the M observations and we obtain a matrix $\mathbf{Z}$, the rows of which contain the new mixtures.

In order to find the deflation parameters given by α_i, we propose the following formulation:

$$\min_{\alpha_i} \quad \|\mathbf{z}_i\|_0, \quad (15)$$

with $\mathbf{z_i}$ given by equation (14).

The resolution of this problem will find a α_i that will give a $\mathbf{z}_i$ with the highest number of zero elements, that is, the found α_i will correct the scale between the original source $\mathbf{s}_i$ and the estimated source $\mathbf{y}_1$, so that $\mathbf{s}_i$ will be removed from $\mathbf{x}_i$.

Again, in order to have a practical approach, we consider a smooth approximation of the l_0-pseudo-norm, so the deflation procedure is based on:

$$\|\mathbf{z}_i\|_0^S = K - \sum_{k=1}^{K} \exp\left(-\frac{z_{ik}^2}{2\sigma^2}\right), \quad (16)$$

where z_{ik} denotes the k-th sample of the new mixture i.

Therefore, the optimization problem to be solved is:

$$\min_{\alpha_i} \quad \left[K - \sum_{k=1}^{K} \exp\left(-\frac{z_{ik}^2}{2\sigma^2}\right)\right]. \quad (17)$$

In this work, this problem was tackled by an exhaustive search procedure, where the limits of our search space were defined as $[-5, 5]$, with step-size of 0.1.

IV. RESULTS

In order to validate the proposed approach, we conducted the following experiments that consists in comparing both methods firstly applied to BSE and after to BSS. The comparison performed for BSE was to analyze the existence of local minima and for BSS was to investigate the obtained signal-to-interference ratios (SIR). Here, it is important to mention that all the parameters were chosen after some numerical simulations, with different values for each one.

A. Comparison between Gradient and Differential Evolution methods in BSE

First, we randomly constructed two sources with uniform probability distribution, in the range $[-0.5, 0.5]$, with 5000 samples each one. In order to let the sources have different sparsities, we randomly made 3500 elements of the first and 400 elements of the second equal to zero. Figure 1 illustrates both sources. Then, these sources are mixed according the following linear mixing model:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

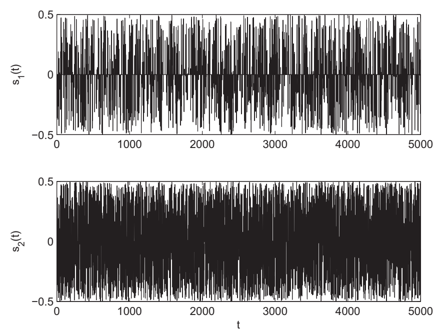

Fig. 1. Sparse sources used for the comparison between the two methods.

Then, we performed the BSE with the gradient descent method and with the differential evolution method. For the analysis by the gradient method, we choose the following parameters: maximum number of iterations = 10000, termination tolerance = 1×10^{-9}, minimum allowed perturbation =

1×10^{-16} and step-size $\mu = 100$. We choose the initial value of $\mathbf{w}$ randomly with uniform probability distribution, in the range $[0, 1]$ and the standard deviation used for the cost function calculation, according to the smoothed l_0-norm model, was $\sigma = 0.01$.

The method stopped with 5800 iterations and, as can be seen in Figure 2, it converged in the considered scenario, because the values of w_1 and w_2 (vector $\mathbf{w}$ components) remained constant since approximately 5600 iterations. Therefore, the method found a local minimum.

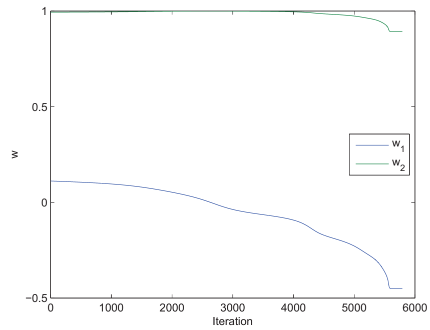

Fig. 2. Evolution of extraction vector elements w_1 and w_2.

For the analysis by the differential evolution method, we used the following parameters: iteration number = 1000, mutation rate $F = 0.5$, crossover rate $CR = 0.5$, population size $NP = 10$. We choose the initial value of all vectors $\mathbf{w}_p$ randomly with uniform probability distribution, in the range $[0, 1]$, and the standard deviation used for the cost function calculation, according to the smoothed l_0-norm model, was $\sigma = 0.01$.

To compare the results obtained by both methods, we constructed Figure 3, where it is possible to see the cost function versus θ, that we used for the parameterization of the problem constraint curve, that was constructed in the following way: $\mathbf{w} = [\cos\theta \ \sin\theta]^T$ with $-\pi \leq \theta \leq \pi$ (this satisfies the problem constraint $\|\mathbf{w}\|_2 = 1$, because $\cos^2\theta + \sin^2\theta = 1$). The black point corresponds to the optimum value given by the gradient method and the red point corresponds to the optimum value given by the differential evolution method. The gradient method converged to a local minimum and the differential evolution method converged to the global minimum. Therefore, since the gradient method is not sufficiently robust to local minima, we can have a convergence to a local minimum, which does not result in a good estimation of the sparsest source.

B. *Performance measures for the methods in BSS*

A way of assessing the efficiency of a BSS method is by the SIR calculation, given by:

$$\text{SIR} = 10 \log \frac{E\{\mathbf{s}^2\}}{E\{(\mathbf{s} - \mathbf{y})^2\}} \tag{19}$$

where $E\{.\}$ is the expected value, $\mathbf{s}$ is the original source and $\mathbf{y}$ is its estimate. So, one has a SIR, given in dB, for

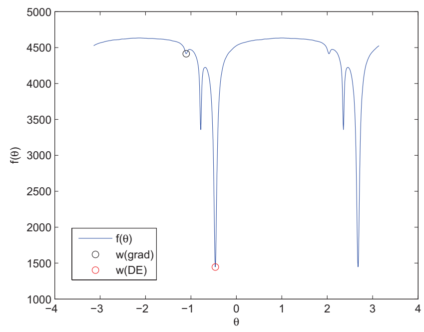

Fig. 3. Cost function versus θ.

each estimated source. It is possible to see that the smaller the difference between $\mathbf{s}$ and $\mathbf{y}$, the bigger the value of the SIR.

So that we could measure the performance of the methods for the BSS, using the gradient and differential evolution methods for the BSE and the exhaustive search for the deflation, we calculated the mean between the mean SIR of 50 simulations for scenarios with two to ten sources, using orthogonal disjoint and non disjoint sources.[1] We randomly constructed the sources with uniform probability distribution, in the range [-0.5; 0.5], with 5000 samples each one. The sparsity, for disjoint sources, was achieved BY making one element different to zero in one source and the corresponding elements in the other sources equal to zero. For the non disjoint sources, we randomly made some elements of each source equal to zero. The sources were mixed according to the linear mixture model and the number of mixtures was the same as the number of sources. We constructed the mixture matrix $\mathbf{A}$ with the elements of its principal diagonal equal to 1 and the other elements randomly generated with uniform probability distribution, in the range $[0.1, 0.9]$. We choose the following parameters for the gradient method: maximum number of iterations = 1000, termination tolerance = 1×10^{-9}, minimum allowed perturbation = 1×10^{-16} and step-size $\mu = 1000$. We choose the initial value $\mathbf{w} = [0 \ 1]$ and the standard deviation used for the cost function calculation, according to the smoothed l_0-norm model, was $\sigma = 0.01$. For the analysis by the differential evolution method, we used the following parameters: iteration number = 1000, mutation rate $F = 0.5$, crossover rate $CR = 0.5$, population size $NP = 10$. We choose the initial value of all vectors $\mathbf{w}_p$ randomly with uniform probability distribution, in the range $[0, 1]$, and the standard deviation used for the cost function calculation, according to the smoothed l_0-norm model, was $\sigma = 0.01$. For the deflation, we choose the following parameters: α_i was searched in the range $[-5, 5]$, with a step-size of 0.01, and the standard deviation used for the cost function calculation, according to the smoothed l_0-norm model, was $\sigma = 0.01$. Figure 4 show the achieved results.

The red and black curves correspond to the differential evolution method, for disjoints and non disjoints sources,

[1]Two sources are orthogonal disjoint if for active elements in one source, the corresponding elements in the other source is inactive, that is, zero.

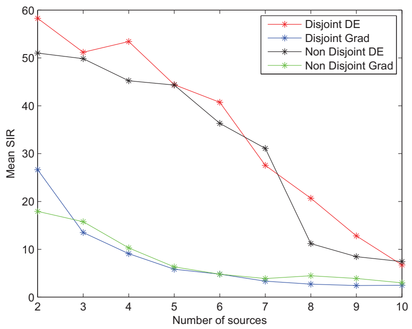

Fig. 4. Mean SIR versus number of sources.

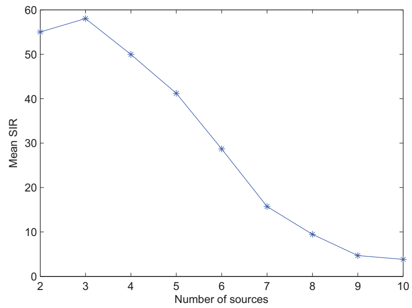

Fig. 5. Mean SIR versus number of sources for the differential evolution methods with disjoint sources.

respectively. The blue and green curves correspond to the gradient method, for disjoint and non disjoint sources, respectively. We verified that the higher efficiency to the BSS was with the differential evolution method, compared with the gradient method. It happened because the initial value was randomly chosen and in some of the simulations with the gradient method, the global minimum was not reached, thus not achieving BSE and decreasing the efficiency of the BSS.

To analyze more accurately the best reached case in BSS, that is, performing the differential evolution method in disjoint sources, we calculated the mean between the mean SIR in 100 simulations for scenarios of two to ten sources. We randomly constructed the sources with uniform probability distribution, in the range [-0.5; 0.5], with 5000 samples each one. We performed the mixtures according to the linear mixture model and the number of mixtures was the same as the number of sources. We constructed the mixture matrix $\mathbf{A}$ with the elements of its principal diagonal equal to 1 and the other elements randomly generated with uniform probability distribution, in the range $[0.1, 0.9]$. For the analysis by the differential evolution method, we used the following parameters: iteration number = 1000, mutation rate $F = 0.5$, crossover rate $CR = 0.5$, population size $NP = 10$. We choose the initial value of all vectors $\mathbf{w}_p$ randomly with uniform probability distribution, in the range $[0, 1]$, and the standard deviation used for the cost function calculation, according to the smoothed l_0-norm model, was $\sigma = 0.005$. For the deflation, we choose the following parameters: α_i was searched in the range $[-5, 5]$, with a step-size of 0.01, and the standard deviation used for the cost function calculation, according to the smoothed l_0-norm model, was $\sigma = 0.01$. Figure 5 show the achieved results.

As expected, there was a decrease of the SIR with the increase of the number of sources due to the higher complexity of the problem, because there is an increase of the search space dimension. However, considering that a SIR at least equal to 12 dB is a satisfactory value, the differential evolution method performance was satisfactory for the separation of two to seven sources, as seen in Figure 5.

V. Conclusions

In this work we developed a method for sparse source separation, that was divided in two parts: blind source extraction and deflation. We showed that the chosen method for blind source extraction, the differential evolution, had satisfactory results in the separation of two to seven sources. When compared to the gradient descent method, the differential evolution method presented better results, because the gradient method converged to local minima, decreasing the efficiency of source separation.

Acknowledgments

H. E. Oliveira thanks FAEPEX (Unicamp) for funding his master research and CAPES for funding his doctoral research. L. T. Duarte thanks FAPESP and CNPq for funding his research.

References

[1] J. M. T. Romano, R. Attux, C. C. Cavalcante, and R. Suyama, *Unsupervised signal processing: channel equalization and source separation*. CRC Press, 2010.

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

[3] Y. Deville, *Blind Source Separation and Blind Mixture Identification Methods*. John Wiley & Sons, Inc., 2016. [Online]. Available: http://dx.doi.org/10.1002/047134608X.W8300

[4] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach," *Signal processing*, vol. 45, no. 1, pp. 59–83, 1995.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2001.

[6] L. T. Duarte, R. Suyama, R. Attux, J. M. Romano, and C. Jutten, "Blind extraction of sparse components based on l_0-norm minimization," in *Statistical Signal Processing Workshop (SSP), 2011 IEEE*. IEEE, 2011, pp. 617–620.

[7] N. Mourad and J. P. Reilly, "Blind extraction of sparse sources," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2666–2669.

[8] E. Z. Nadalin, A. K. Takahata, L. T. Duarte, R. Suyama, and R. Attux, "Blind extraction of the sparsest component," in *Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 394–401.

[9] Y.-M. Batany, D. Donno, L. T. Duarte, H. Chauris, Y. Deville, and J. M. T. Romano, "A necessary and sufficient condition for the blind extraction of the sparsest source in convolutive mixtures," *EUSIPCO*, 2016.

[10] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[11] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed norm," *Signal Processing, IEEE Transactions on*, vol. 57, no. 1, pp. 289–301, 2009.