Sequential Sparse Blind Source Separation for Non-Linear Mixtures

C Kervazo\textsuperscript{1}, J Bobin\textsuperscript{1}
\textsuperscript{1} CEA Saclay, Gif-sur-Yvette, 91191 cedex, France
E-mail: christophe.kervazo@cea.fr

Abstract. Linear Blind Source Separation (BSS) has known a tremendous success in fields ranging from biomedical imaging to astrophysics. In this work, we however propose to depart from the usual linear setting and tackle the case in which the sources are mixed by an unknown non-linear function. We propose to use a sequential decomposition of the data enabling its approximation by a linear-by-part function. Beyond separating the sources, the introduced StackedAMCA can further empirically learn in some settings an approximation of the inverse of the unknown non-linear mixing, enabling to reconstruct the sources despite a severely ill-posed problem. The quality of the method is demonstrated experimentally, and a comparison is performed with state-of-the art non-linear BSS algorithms.

1. Linear and Non-Linear BSS

1.1. Context

Since its formulation in the 1980s, Blind Source Separation (BSS) has become one of the major tools to learn meaningful decompositions of multivalued data. It has been used in many fields, such as audio processing, biomedical imaging or astrophysics [1–3].

Most of this work has however been dedicated to linear BSS, in which \( m \) observations are assumed to be the linear combinations of \( n \) sources, each of them having \( t \) samples. In matrix form, it is supposed that the data can be written as \( X = AS + N \), with \( X \) (size \( m \times t \)) the observation matrix corrupted with some unknown noise \( N \). The sources \( S \) (\( n \times t \)) are therefore supposed to be mixed linearly through the \( A \) matrix (\( m \times n \)) and the goal of linear BSS is to recover both \( A \) and \( S \) from the sole knowledge of \( X \) up to a permutation and scaling indeterminacy. While linear BSS is ill-posed, several types of priors have been introduced to reduce the space of possible solutions. Among them, sparsity [4] – which assumes that the sources have a large number of zero coefficients – has been shown to lead to enhanced separation quality on various problems of linear BSS [5–7].

While convenient for many problems, the linear mixing model is only an approximation which might not hold in various experimental settings. For instance, it is not anymore valid when using sensors with saturations or non-linearities (for instance gas [8] or chemical [9] sensors), or in some specific applications (show-through removal [10], hyperspectral imaging [11]). In all these applications, it is therefore relevant to change the BSS model to a non-linear one:

\[
X = f(S) + N
\]  
(1)

Where \( f \) is an unknown non-linear function from \( \mathbb{R}^{n \times t} \) to \( \mathbb{R}^{m \times t} \) (here, \( n \leq m \)). In this work, we will consider general functions \( f \), by mostly (cf. Sec. 4) assuming that \( f \) is invertible and
symmetrical around the origin, as well as regular enough (cf. Sec. 4). In particular, such a regularity means that \( f \) must not deviate from a linear mixing too fast as a function of the input amplitude: this can for instance be satisfied on practical experiments with chemical sensors [12] or in which sensor saturations can occur.

At this point, it is important to mention that non-linear BSS is much more difficult than its linear counterpart and that it might not be possible to find both \( f \) and \( S \) up to a simple permutation and scaling indeterminacy. In the case of sparse sources, [13, 14] have however shown the possibility to recover the sources \textit{up to a nonlinear function} if only one source is active for each sample. Therefore, the problem is too ill-posed to ensure a good reconstruction of the sources in the general case, and the goal of sparse non-linear BSS is only to separate the sources by estimating the underlying non-linearities. Beyond the use of sparsity (for which the current algorithms mainly use a clustering approach, which is very different from ours), the most common approach is to use as prior the statistical independance of the sources (Independant Component Analysis – ICA – methods). Contrary to the linear case, this prior is nevertheless not anymore sufficient to separate the sources in the general non-linear setting [15], and therefore most authors use stronger assumptions [16, 17].

1.2. Contribution
We propose to tackle the general problem of \textit{non-linear} BSS presented in Eq. (1) by using a sparsity prior on the sources. To the best of our knowledge, our method is the first attempting to find a linear-by-part approximation of the underlying non-linearities using a sequential sparse BSS approach. Beyond separating them, the algorithm proposes a possible reconstruction of the sources by inverting the estimated linear-by-part model. Despite the usual non-linear BSS indeterminacies, the proposed reconstruction is empirically shown to estimate well the true sources in some settings. In Sec. 2, the method is further described. In Sec. 3, our method is experimentally compared to other ones to show its relevance. In Sec. 4, the required hypotheses of the proposed approach are studied.

1.3. Notations
Scalars are denoted as lower case letters (e.g. \( \tau \)), matrices in bold upper case letters (e.g. \( \mathbf{X} \)), and their estimation by an algorithm as \( \hat{\mathbf{X}} \). The notation \( \mathbf{X}_r \) is the row vector corresponding to the \( r \)th row of the matrix \( \mathbf{X} \) (\( \mathbf{X}_{1..r} \) being the set of rows from index 1 to \( r \)), while \( \mathbf{X}^i \) is the specific sample vector of \( \mathbb{R}^m \) indexed by \( i \). The \( \mathbf{X}^\dagger \) symbol stands for the Moore-Penrose pseudo-inverse of \( \mathbf{X} \). Functions with matrix outputs are written as \( f \). In iterative algorithms, the estimate of a variable \( a \) at the \( l \)th iteration is denoted as \( a^{(l)} \). The set of all the variables estimated between the iterations 1 to \( l \) is denoted as \( a^{(1..l)} \).

2. Proposed Approach
2.1. A Geometrical Perspective on Sparse Non-Linear BSS
The proposed method is described by adopting a geometrical point of view in the case \( n = 2 \); the generalization of the principle to higher values is straightforward.

Due to the morphological diversity assumption [18], it is very rare that sparse sources both have non-zeros values at the same time. Therefore, when plotting the scatter plot of \( \mathbf{S}_1 \) as a function of \( \mathbf{S}_2 \) (cf. Fig. 1), most of the source coefficients lie on the axes (in this work we even assume that all coefficients lie on the axes – this hypothesis is discussed in Sec. 4). Once mixed with the \textit{non-linear} \( f \), the source coefficients lying on the axes are transformed into \( n \) non-linear one dimensional (1D) manifolds [13, 14], each manifold corresponding to one source (see Fig 1). To separate the sources, the idea is then to back-project each manifold on one of the axes. We propose to perform this back-projection by approximating the 1D-maneifolds by a \textit{linear-by-part} function, that we will invert. As evoked above and due to the source reconstruction
indeterminacy, we then get separated sources which are only distorted through an unknown non-linear function that does not remix them, called \( h \) in the following.

2.2. Overview of the Proposed Approach

As can be seen in Fig. 1, the lowest amplitude data coefficients can be well approximated by a classical linear model because of the regularity assumption on \( f \), stating that the data must not deviate from linearity too fast as a function of the amplitude. Finding such an approximation can in practice be done using a sparse linear BSS algorithm, provided that this one is robust to the higher amplitude non-linearities. A rough estimate of the sources can then be computed by inverting the found linear model. As expected, the corresponding separation is however very poor for the higher amplitude highly non-linear samples, as seen in Fig. 3, where such samples (outside of the red square) do not lie at all on the axes as we would like. The question is then: how to better separate these samples? This is done by introducing a non-linear selection step enabling to remove the contribution of the previously found linear model, creating a new dataset \( \mathbf{R} \) comprehending only the highest non-linear samples, which amplitudes are further shrinked. Since the amplitudes are then smaller, working on \( \mathbf{R} \) (cf. Fig. 3) makes possible the estimation of a new linear model that separates better the originally higher amplitude coefficients. It is then possible to repeat the procedure to improve the separation of still higher non-linear samples. The whole procedure, as well as some notations that will be developed in the following detailed explanation of the two main steps, are summarized in Fig. 2.

2.3. Detailed Description

2.3.1. Linear Sparse BSS Step: AMCA

The linear sparse BSS algorithm must find a linear model representing well the lowest amplitude samples of the residual \( \mathbf{R} \), while being insensitive
Figure 3. Illustration of the main steps of the algorithm on the non-linear mixing of Fig. 1. 
Upper left: in blue, output of linear BSS step: $\hat{S}^{(1)}$ is displayed. Compared to Fig. 1, inverting the linear model corresponds to align the found dashed arrows of Fig. 1 with the axes. In addition, the red square delimits the low amplitude sample areas of of $S^{(1)}$ where the linear model is a good approximation – the corresponding maximum amplitudes are denoted by $\tau^{(1)}_{1..n}$ – which means the areas where the points almost lie on the axes; Down right: Residual $R$ after the selection step. In brief, removing the contribution of the found linear model is done by shrinking the amplitudes of the samples in $X$ by $\tau^{(1)}_{1..n}$.

to the higher amplitude samples that are more affected by the non-linearities. We propose to use the AMCA algorithm [6], which is taylored to separate sources having large amplitudes coefficients with partial correlations (i.e. multiple sources are simultaneously active). In brief, AMCA discards the highly non-linear samples by considering them as partial correlations. It is thus able to find at iteration $l$ a good linear model $\hat{A}^{(l)}$ of the lowest amplitude samples of $R$, which once inverted aligns the lowest amplitude samples of each 1D-manifold with the axes. The result is denoted as: $\hat{S}^{(l)} = \hat{A}^{(l)} R$ (cf. Fig. 3).

2.3.2. Selection Function: Computing $R$ The goal of the selection function is to extract within the sources $\hat{S}^{(l)}$ the contributions that are not explained by the linear model found with the BSS step. For finding such contributions, there are two issues: i) **determine which samples are well separated by the current linear model $\hat{A}^{(l)}$**; ii) **actually compute $R$ by shrinking $\hat{S}^{(l)}$ to remove the contribution explained by the current linear model**.

Solution to problem i) StackedAMCA uses for each sample of $\hat{S}^{(l)}$ the distance to the axes. If such a distance is small enough, it is assumed that the sample is well separated by the current linear model. Then, for each source $k$, the threshold $\tau^{(l)}_k$ required for the amplitude shrinkage (see below) is roughly chosen as the maximum amplitude of the samples close enough to the axis of $k$ (the maximum distance depending on the manifold regularity and being an hyper-parameter of the algorithm). In practice, this technique is enhanced with a clustering method to enhance the robustness.

Solution to problem ii) Once we have found the thresholds $\tau^{(l)}_{1..n}$, $\hat{S}^{(l)}$ must be shrunk correspondingly, which is not trivial. In brief, to attribute the highly non-linear samples to the corresponding source we resort to a first guess based on the distance with each axis, thus determining in which direction the shrinking must be performed. This first guess is enhanced starting back from the raw data $X$ and unrolling the manifolds using the linear models found in the previous iterations, both enabling a better constrast between the sources and decreasing error propagation.

3. Experiments
In this part, we present an experiment attesting the quality of the proposed method.
3.1. Metric
The metric choice for non-linear BSS is made more complicated due to the indeterminacy by the non-linear function \( h \). It is thus important to differentiate measures about the reconstruction of the sources and about their separation.

3.1.1. Metric for the Separation Quality
A classical approach to determine the separation quality is to estimate \( h \) [19] by fitting a non-linear curve \( \mathcal{P} \) to the 1D-manifold of the scatter plot of each estimated source \( \hat{\mathbf{S}}_r \) as a function of the true one \( \mathbf{S}_r \) and to look at the thickness of the manifold around \( \mathcal{P} \). Here, we use for \( \mathcal{P} \) a polynomial function of degree 20 and the thickness is measured by:

- The logarithmic median distance with the curve \( \mathcal{P} \): \( C_{\text{med}} = -10 \log(\sum_{r=1}^{n}\text{median}_{i}(\hat{S}_r^i - \mathcal{P}(\mathbf{S}_r^i))) \)
- The logarithmic Euclidian distance with the curve \( \mathcal{P} \): \( C_{\text{sq}} = -10 \log(\sum_{r=1}^{n}\frac{1}{t}\sqrt{\sum_{i=1}^{t}(\hat{S}_r^i - \mathcal{P}(\mathbf{S}_r^i))^2}) \)

However, the results of these metrics are sensitive to the choice of \( \mathcal{P} \). We thus propose to introduce for exactly sparse sources a new metric based on the angular distance to the axes:

\[
C_{\text{ang}} = -10 \log \left( \frac{1}{n(n-1)} \sum_{r=1}^{n} \left( \sum_{i \neq j}^{n} 1 - \frac{1}{\#Z} \sum_{t \in Z} \frac{\mathbf{S}_r^i}{\sqrt{\mathbf{S}_r^i \cdot \mathbf{S}_r^j}} \right) \right)
\]

(2)

where \( Z = \{ t|\mathbf{S}_r^t \neq 0 \} \) and \( \#Z \) denotes the cardinal of \( Z \).

3.1.2. Metrics for Source Reconstruction
To determine whether the source reconstruction is good or not, it is possible to use classical metrics between the estimated and true sources, such as the Mean Squared Error (MSE) and the Mean Error (ME).

3.2. Experimental Results and Comparison to Other Methods
In this section, we compare the results of our algorithm to other existing ones. Only a few algorithms for non-linear BSS are open source, and we mostly found three of them: MISEP [20], NFA [21] and ANICA [22]. The experiment itself comes from [13]. The sources follow a Bernouilli-Gaussian ditribution, \( p = 10\% \) of the \( t = 9500 \) samples being non-zeros. The supports of the \( n = 2 \) sources are disjoint (experiments with more sources have also been performed, leading to the same conclusions). There is \( m = n = 2 \) observations, which are computed for each element indexed by \( i \in [1, t] \) as \( \mathbf{X}_i = \cos(\alpha(i))\mathbf{S}_i^1 - \sin(\alpha(i))\mathbf{S}_i^2 + \mathbf{N}_i^1 \) and \( \mathbf{X}_i^2 = \sin(\alpha(i))\mathbf{S}_i^1 + \cos(\alpha(i))\mathbf{S}_i^2 + \mathbf{N}_i^2 \) with \( \alpha(i) = \frac{\pi}{2}(1 - \sqrt{\mathbf{S}_i^1 \cdot \mathbf{S}_i^2}) \) and \( \mathbf{N} \) chosen such that the SNR is 30 dB.

The different separation and reconstruction metrics are displayed in Table 1. The corresponding results are shown in Fig. 4, where the scatter plot of one estimated source is drawn as a function of the true one. First, it seems that neither ANICA nor NFA truly separate the sources (concerning ANICA, the results seem however to improve when no noise is added). Although it could come from our lack of familiarity with the parameter tuning of these methods, it is possible that the regularization introduced by the network structure for ANICA and the Bayesian setting for NFA is not sufficient to enable the separation of the sources (since the independance is not either, cf. Sec 1). On the contrary, MISEP separates the sources well. The separation performed by StackedAMCA is however better than with all the other methods, especially when using metrics that are robust to a small number of outliers.

Second, MISEP does not reconstruct well the sources as StackedAMCA does and Fig. 4 clearly indicates that it did not invert the non-linearity \( h \). On the contrary, the good ME
Table 1. Separation and reconstruction quality of StackedAMCA, MISEP, NFA and ANICA.

| Method      | $C_{med}$ | $C_{sq}$ | $C_{ang}$ | MSE  | ME   |
|-------------|-----------|----------|-----------|------|------|
| StackedAMCA | 49.7      | 46.5     | 36.7      | 43.6 | 30.2 |
| MISEP       | 26.7      | 44.8     | 18.3      | 34.3 | 20.7 |
| NFA         | 16.7      | 30.9     | 4.09      | 30.1 | 14.1 |
| ANICA       | 19.9      | 34.4     | 1.56      | 19.4 | −0.206 |

Figure 4. Scatter plot of one estimated source as a function of the true source. From left to right, results of: StackedAMCA, MISEP, NFA, ANICA.

of StackedAMCA indicates that the algorithm structure was sufficient to regularize well the reconstruction problem. As a side remark, it is interesting to note that a class of non-linearities $f$ for which StackedAMCA is able to perform such a good reconstruction can be characterized.

4. Required Hypotheses for StackedAMCA

4.1. Symmetry of $f$ Around the Origin
The symmetry assumption could in principle be leveraged. First, the data can be symmetrical around a different point as long as a preprocessing step is introduced to center it. Then, tackling non-symmetrical data could probably be dealt with by introducing non symmetrical non-linear steps and adding a non-negativity constraint in the linear BSS step.

4.2. Disjoint Supports
We have assumed the supports of the sources to be disjoint. While this is not very realistic in practical cases, it seems difficult to bypass this condition as we only explore the span of $f$ that the 1D-manifolds created by the sparse sources uncover. By the morphological diversity assumption, the points outside these manifolds are too rare to enable a proper estimation of $f$ without any further conditions (e.g. the separability over the different sources). We however emphasize that we did some tests without disjoint supports. In this case, the samples with multiple active sources were badly separated but the estimation of the 1D-manifolds by StackedAMCA was not much perturbated, which is mainly due to the robustness of AMCA to multiple active sources.

4.3. Regularity of the Mixing $f$
We have assumed that $f$ does not deviate too fast from linearity as a function of the amplitude. For differentiable curves, it mathematically means that at every point of the 1D-manifolds described by the mixing $X$, the local curvature radius must be large enough. This condition is of primary importance to enable StackAMCA to separate the sources and leverage the issue of potential permutations between layers (which is also alleviated thanks to a permutation
correction, not described here for the sake of simplicity). For a similar reason, $f$ must also be $L$-Lipschitz with $L$ small enough.

5. Conclusion
We introduce in this work StackedAMCA, a new algorithm tackling the sparse non-linear BSS problem by sequentially computing a linear-by-part approximation of the underlying non-linearities. Each linear part is estimated by a robust linear BSS algorithm step, which is followed by a selection step. The non-linear selection step enables to work on increasingly higher non-linearities. We show the relevance of StackedAMCA compared to other state-of-art methods. Beyond separating the sources, in some experiments the algorithm is also able to reconstruct them well despite a severely ill-posed problem. A discussion of the required hypotheses for StackedAMCA to work is furthermore proposed.

Acknowledgment
This work is supported by the European Community through the grant LENA (ERC StG - contract no. 678282).

References
[1] Vincent E, Jafari M G, Abdallah S A, Plumbley M D and Davies M E 2011 Probabilistic modeling paradigms for audio source separation Machine Audition: Principles, Algorithms and Systems (IGI Global) pp 162–185
[2] Negro F, Muceli S, Castronovo A M, Holobar A and Farina D 2016 Journal of neural engineering 13 026027
[3] Bobin J, Sureau F, Starck J L, Rassat A and Paykari P 2014 Astronomy & Astrophysics 563 A105
[4] Zibulevsky M and Pearlmutter B A 2001 Neural Computation 13 863–882
[5] Bobin J, Starck J L, Moudden Y and Fadili M J 2008 Advances in Imaging and Electron Physics 152 221–302
[6] Bobin J, Rapin J, Larue A and Starck J L 2015 IEEE Transactions on Signal Processing 63 1199–1213
[7] Kervazo C, Bobin J and Chenot C 2018 Signal Processing 150 157–165
[8] Madrolle S, Grangeat P and Jutten C 2018 Sensors 18 1785
[9] Jimenez G B 2006 Non-linear blind signal separation for chemical solid-state sensor arrays Ph.D. thesis Universitat Politècnica de Catalunya
[10] Merrikh-Bayat F, Babaie-Zadeh M and Jutten C 2011 International Journal on Document Analysis and Recognition (IJDAR) 14 319–333
[11] Dobigeon N, Tourneret J Y, Richard C, Bermudez J C M, McLaughlin S and Hero A O 2014 IEEE Signal Processing Magazine 31 82–94
[12] Duarte L T and Jutten C 2014 Oil & Gas Science and Technology–Revue d’IFP Energies nouvelles 69 293–306
[13] Ehsandoust B, Rivet B, Jutten C and Babaie-Zadeh M 2016 Nonlinear blind source separation for sparse sources Signal Processing Conference (EUSIPCO), 2016 24th European (IEEE) pp 1583–1587
[14] Puigt M, Griffin A and Mouchtaris A 2012 Nonlinear blind mixture identification using local source sparsity and functional data clustering Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th (IEEE) pp 481–484
[15] Comon P and Jutten C 2010 Handbook of Blind Source Separation: Independent component analysis and applications (Academic Press)
[16] Deville Y and Duarte L T 2015 An overview of blind source separation methods for linear-quadratic and post-nonlinear mixtures International Conference on Latent Variable Analysis and Signal Separation (Springer) pp 155–167
[17] Hyvarinen A and Morioka H 2017 Nonlinear ICA of temporally dependent stationary sources (Proceedings of Machine Learning Research)
[18] Bobin J, Starck J L, Fadili J M and Moudden Y 2007 IEEE Transactions on Image Processing 16 2662–2674
[19] Ehsandoust B, Babaie-Zadeh M, Rivet B and Jutten C 2017 IEEE Transactions on Signal Processing 65 4339–4352
[20] Almeida L B 2003 Journal of Machine Learning Research 4 1297–1318
[21] Honkela A, Valpola H, Ilin A and Karhunen J 2007 Digital Signal Processing 17 914–934
[22] Brakel P and Bengio Y 2017 arXiv preprint arXiv:1710.05050