BLIND SOURCE SEPARATION OF ASTRONOMICAL IMAGES

Danielle Nuzillard

Université de Reims Champagne-Ardenne
L.A.M., Moulin de la Housse
BP 1039, 51687 Reims Cedex 2, France
Danielle.Nuzillard@univ-reims.fr

Albert Bijaoui

Observatoire de la Côte d'Azur
Dpt CERGA B.P.4229
06304, NICE CEDEX 4, France
Bijaoui@obs-nice.fr

ABSTRACT

Celestial sources are studied by means of multispectral analysis, for which a pixel value can be considered as a mixture of different sources. We show that the application of blind source separation (BSS) methods can display interesting features, which could help the astronomers to improve their description of celestial sources. Experiments were done on Hubble Space Telescope (HST) images of the galaxy 3C120. Different ways were applied for the separation: tools based on the correlation between the shifted sources, and tools based on non-Gaussian probability density functions (PDF). Each obtained source set was quantified by the mutual information between the sources. Different algorithms give quite similar decompositions which mainly correspond to real phenomena.

1. MULTISPECTRAL ANALYSIS OF ASTRONONOMICAL IMAGES.

The spectral distribution of celestial sources carries out essential information on the physical processes which take place in these objects. Therefore, since many decades the astronomers have developed multispectral observations first by means of photographic plates and now using CCD detectors. Each image is taken throughout a colored filter which delimits a well-defined spectral band. From the flux measurements for each filter, the astronomers compute color indexes from which they deduce source temperatures, stellar gravities, or non-thermal properties.

The color indexes allow the astronomers to classify the pixels. Pixel classification suffers of a drawback, the mixture: each pixel value can be considered as the result of a mixing between different sources. Then an available classification needs to decompose the mixtures into independent sources.

The spectral energy of a celestial objet is distributed among a continuum and spectral lines which are superimposed. The spatial distribution is determined by the transfer radiation equation which is linked to non linear phenomena. Consequently, the astronomers can not claim that the colored sky images may be written as a weighted combination of independent sources.

Nevertheless, we will show in this communication that the application of BSS methods can display interesting features, which could help the astronomers to improve their description of celestial sources.

2. THE DATA SET.

2.1. The observations.

The radiogalaxy 3C120 is a curious celestial source. It displays a one side radio jet, which has been optically identified [6] and where superluminal motions have been detected. This object is interpreted as the consequence of the merging of two galaxies. At the center of the main component a black-hole swallows down the matter from the second component. Two jets are emitted from the black hole, the visible one coming to us near the line-of-sight.

3C120 was observed at different periods with HST, the best observations were done with the *Wide Field Planetary Camera 2* (WFPC2) throughout filters F547M (a filter centered at the wavelength 547 nm with a medium width), F555W (555 nm with a wide width), F675W (675 nm with a wide width), and F814W (814 nm with a wide width).

F547M contains two strong twice ionized oxygen lines [OIII]. F555W recovers F547M, the other part contains only the continuum. F675W contains the strong ionized hydrogen [HII] H_α line. F814W contains [OI] and [SIII] lines. The combinations of the images display specific physical phenomena. Then we expect from BSS to obtain images which show the best distribution of these phenomena.

2.2. The data preparation.

The central WFPC2 observations correspond to images of 800×800 pixels. The pixel size is 0.0455" in two dimensions, and the resolution is around 0.1". We have extracted the central part of 256×256 pixels, corresponding to a region of $11''.6 \times 11''.6$.

The observations were done with a CCD receiver. The detector noise is low, and the photon noise is dominant in the galaxy region. As the noise is not stationary, we applied a generalized Anscombe's transform [8] which allows us to stabilize the variance. For each resulting image, the background value is estimated and subtracted. Then the images (figure 1) can be considered clean enough to be processed by BSS algorithms.

3. BSS EXPERIMENTS.

Since many decades Karhunen-Loève expansion (KLE) is applied for extracting the main information from a set of celestial images [1], [7]. In the case of a Gaussian distribution, KLE is sufficient, but generally the distributions are not Gaussian, and a null covariance does not mean that the sources are independent. The resulting demixed images show clearly that non independence still exists (figure 2). KLE allows us to whiten the images and it is considered as the first step of ICA.

We have experimented different ways for improving the separation using BSS algorithms. They consist in recovering a set of n independent signals from m observed mixtures where $m \geq n$. The m observed data are written as a linear combination of n sources.

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

where X is the observed data, A the mixing, S the sources and N the noise matrices. Thanks to the generalized Anscombe's transform, we can assume that the noise is white.

SOBI [2] is an efficient second order algorithm. It lies on the hypothesis of temporally correlated sources and allows an efficient second order separation. It is based on the wish to decrease the cross correlation between shifted sources taken two by two, and it increases the correlation between themselves. We compared the results to those obtained with JADE [3], and with FastICA [5] which are based on search of non-Gaussian probability density functions (PDF) of the sources.

3.1. Comparison methods.

We have done many experiments with these algorithms. For each one we get a mixing and a demixing matrices and image sources. Even so a visual test is quite informative, tools for displaying these results and for getting a quantitative quality measurement are needed.

As we have four images, we limit the number of independent sources also to four. For displaying the demixing matrix, each line is normalized to be unitary. The three first components $(F1, F2, F3)$ are plotted by their two projections in planes $(F1, F2)$ and $(F1, F3)$ then these projections are linked together. For each BSS a demixing diagram of four segments is then plotted.

The source images are displayed in order to optimize the contrast. That allows us to compare the different sources in the best contrast conditions. This visual comparison was essential to select the best identification but it is too qualitative.

A source set was then quantified using a mutual information between the sources. That was done in three steps:

1. For each source i we determine the mean value m_i and the standard deviation σ_i after a $3-\sigma_i$ clipping. In this algorithm we compute iteratively these parameters and we reject the values outside the interval $[m_i - 3\sigma_i, m_i + 3\sigma_i]$. After few iterations $(4-5)$ the algorithm converges. In case of a true Gaussian distribution, the obtained mean is right, while the σ bias is order of 2%. On the contrary, these parameters define a Gaussian kernel of the PDF, and we will measure more values outside the interval $[m - 3\sigma, m + 3\sigma]$ than for a Gaussian PDF.

2. The histogram $H_i(k)$ of source i is determined with a box size equal to this deviation σ_i. We evaluate the entropy of the source E_i by:

$$E_i = -\sum_k \frac{H_i(k)}{N} \log_2 H_i(k) + \log_2 N \tag{2}$$

   where N is the number of pixels. This entropy depends on the box size. If the box is too large, E_i is too small, and it is not sensitive to non-Gaussian features, if the box is too small the number of pixels per box is too low and then the estimation is not available.

3. We determine then the mutual histogram H_{ij} with the same box size, and we compute the resulting entropy E_{ij}. The mutual information between i and j is equal to:

$$I_{ij} = E_i + E_j - E_{ij} \tag{3}$$

   This mutual information is quite independent of the box size, for values smaller than σ_i.

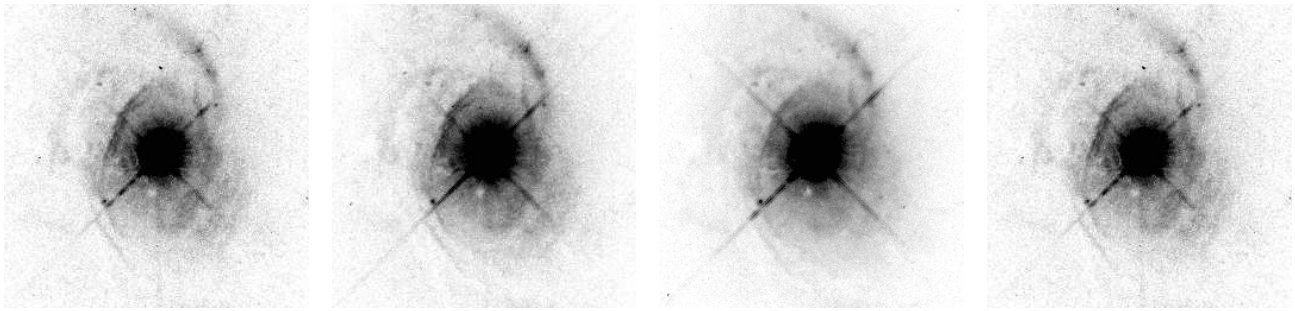

Figure 1: HST/WFPC2 images of the radiogalaxy 3C120 respectively obtained with the filters F547M, F555W, F675W and F814W. The images were processed in order to stabilize the noise variance.

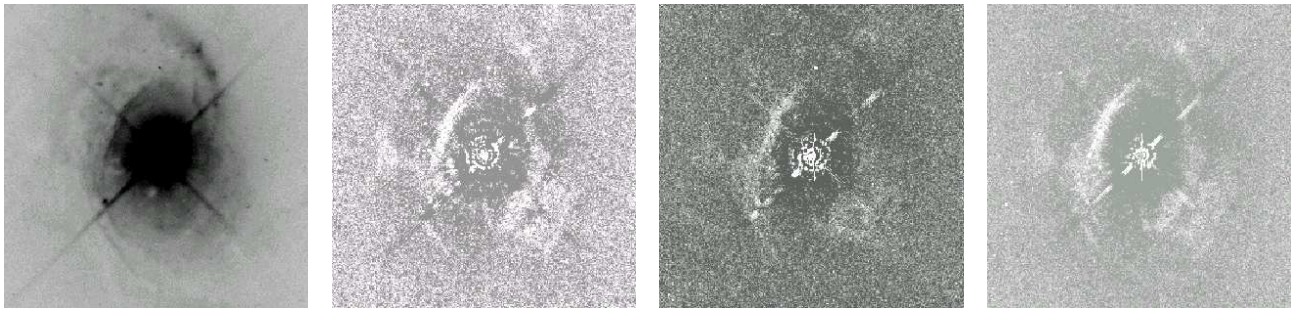

Figure 2: The Karhunen-Loève expansion of the 3C120 HST images.

4. We quantify each demixture by the sum:

$$I = \sum_{i<j} E_{ij}. \tag{4}$$

We note that I only takes into account the pixel PDF. Then it favors the algorithms based on PDF, such as JADE and FastICA.

3.2. SOBI experiments.

SOBI [2] is an algorithm based on the reduction of the cross correlation between the sources and the shifted sources. The number of shifts p is a free parameter. The shifts s_i, $i \in (1, p)$, are free parameters. After the data whitening, a set of p covariance matrices is computed. Cross correlation terms are minimized, thus diagonal terms are maximized. A joint diagonalization criterium of several p covariance matrices improves its robustness [4]. We have adapted the algorithm from the temporal field to the $2D$ one.

In its original version we call it SOBI1, we compute the cross correlation matrices in the direct space. In our application the pixels are aligned, and the algorithm is applied.

For analyzing Nuclear Magnetic Resonance (NMR) spectrograms, Nuzillard [9] has modified SOBI by computing the cross correlations of the Fourier transform, which are easily estimated in the direct space. This can be viewed as an alternative correlation choice. SOBI1 takes into account the correlation at short distances, while the correlations at short frequency distances play the main role for SOBI2. NMR spectrograms displaying narrow peaks (or lines) , so that the correlation rapily decreases, then SOBI2 brings a better separation. For present data this argument is not so clear. The cross correlation between two vectors A and B is expressed as:

SOBI1 In the direct space:

$$R_{AB}(\xi) = \frac{1}{L} \sum_{k} a_k . b^*_{k-\xi} \tag{5}$$

where ξ is the spatial shift and L the number of samples, a_k and b_k are elements of A and B, the sampling step being 1.

SOBI2 In the Fourier space:

$$R_{AB}(n) = \sum_k a_k . b_k^* . e^{-2i\pi(\frac{kn}{L})} \quad (6)$$

where n is the frequency shift.

SOBI3 and SOBI4 are respectively SOBI2 and SOBI1 adaptations to $2D$ images. The algorithms take into account the correlation matrices between two images in different directions: lines, columns, 45°, ...

SOBI3 In the Fourier transform space:

$$R_{AB}(n,m) = \sum_k \sum_l a_{kl} . b_{kl}^* . e^{-2i\pi(\frac{nk}{L_1} + \frac{ml}{L_2})} \quad (7)$$

where n and m are the spatial frequency shifts, L_1 and L_2 the row and column numbers and a_{kl} and b_{kl} pixel values of images A and B.

SOBI4 In the direct space:

$$R_{AB}(\xi,\eta) = \frac{1}{L_1 L_2} \sum_k \sum_l a_{kl} . b_{k-\xi,l-\eta}^* \quad (8)$$

where ξ and η are the spatial shifts.

These four algorithms, with different numbers of cross correlation matrices p, eventually chosen on different directions have been tested. According to the mutual information, SOBI2 with 16 matrices gives the best solution with $I_{16} = 0.581$ bit.

In figure 3 we have given its demixing diagram. In figure 4 the corresponding sources are plotted. Each source displays interesting features:

Source 1 It corresponds clearly to a set of rings around the nucleus. These rings could be also seen in the KL expansion, but SOBI2 displays a cleaner vision. Taking into account their size, these rings do not correspond to Airy's pattern. We have compared this image to the point spread function (PSF) of the WFPC2 at 675 nm obtained with the program TINY (figure 5). Even if the contrasts are not the same, we can assert that the rings correspond to that function. The demixing coefficients show that this source corresponds mainly to the H_α emission line of the center of the radiogalaxy.

It is amazing that this BSS allows us to display a real physical phenomenon, which is due to the fact that a large part of the energy in image

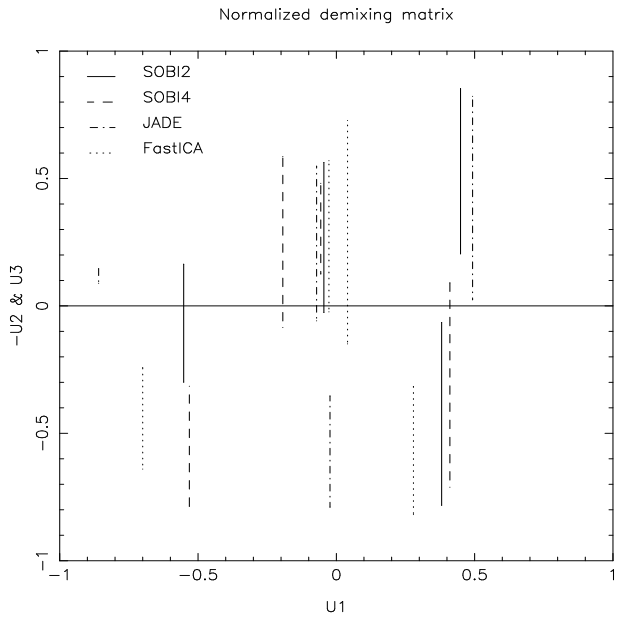

Figure 3: The demixing diagrams obtained for four different algorithms.

F675W is emitted in a spectral line in a very small region, smaller than the diffraction limit. For this line we have the image of the PSF, which displays rings due to the central occultation of the telescope. This source has been separated because the size of the continuum region (due to the jet) has a larger size.

Source 2 It is the less informative source. It looks like to a smooth version of the source 3.

Source 3 It is the more interesting source for a physical insight. It corresponds to the ionized regions surrounding the galaxies. These features carry out an important information on the merging process.

Source 4 This source carries out the maximum entropy and displays the central region. We note the imprint of the source 3.

SOBI2 has then extracted three sources which seem to be correspond to three independent physical components. Some imprints still exist but they are faint compared to the original images. The source 2 may be interpreted as the residue of the separation. Non linearities would play an important part.

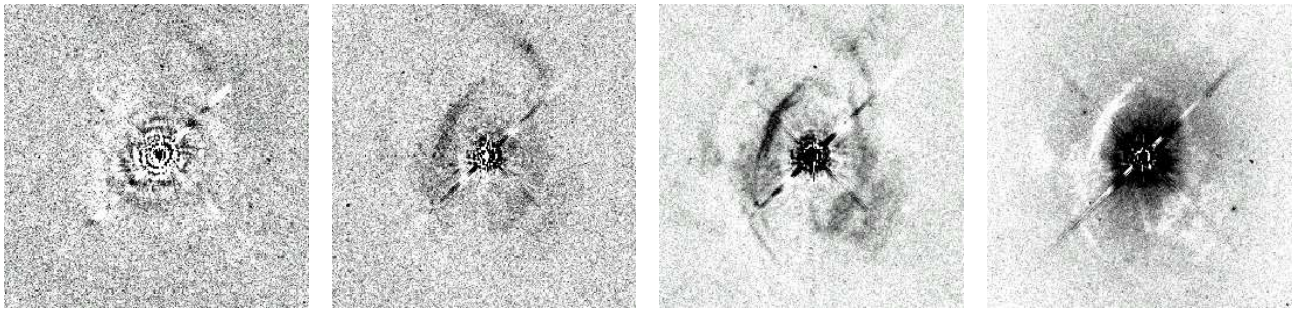

Figure 4: BSS using SOBI2 algorithm with 16 cross correlation matrices.

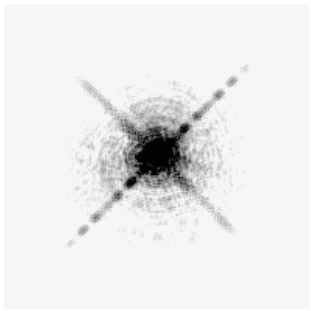

Figure 5: The point spread function at the wavelength 675 nm of the WFPC2 camera.

A visual inspection leads to select another set of sources, the one carried out by SOBI4 with 8 matrices. In figure 3 its demixing diagram is drawn. We find again the sources 1, 3 and 4, but the source 2 contains fewer information (1.55 bit), and it corresponds quite only to the central part. Its mutual information is 0.805 bit (figure 6). The main part corresponds to a large mutual information between source 3 and 4, which is contrary to the visual inspection. I computation takes only into account the pixel values at the same position, while in the visual inspection we could take into account the surrounding pixels. A more sophisticated criterion, taking into account close pixels, has to be developed.

3.3. Comparison with JADE et FastICA

With these algorithms the separation is based on non-Gaussianities of the sources PDF. JADE and FastICA maximize a contrast function. JADE is based on the cumulant order 4. With this algorithm we get a set of sources whose the mutual information is 0.621 bit, greater than I_{16}. The sources look like to the ones obtained with SOBI2 and 16 matrices. The demixing matrix is drawn in figure 3.

With FastICA [5] non gaussianity is measured by fixed-point algorithm using an approximation of negentropy. The nonlinearity function $g(u)$ can be u^3, $\tanh(u)$ or $ue^{-\frac{u^2}{2}}$.

Six tests were done with FastICA. The best solution was obtained with a deflation algorithm where $g(u)$ was $\tanh(u)$ with a mutual information equal to 0.584, very similar to I_{16} (figure 7). The demixing coefficients are also drawn in figure 3.

We can note that a second order blind identification algorithm, based on correlations in a large region around a pixel, carries out the same decomposition than a blind identification related to local high order statistics. That fact brings some confidence to the resulting decomposition.

4. CONCLUSION

In the present communication we have examined the multispectral analysis as a peculiar BSS case. Even if the *cocktail party* model is not realistic for processing celestial images, we have shown that the application of different algorithms gives quite similar decomposition.

From a physical insight the resulting sources mainly corresponds to real phenomena. Then these separations carry out a useful information for the astrophysicists.

For the best separation selection we have introduced a simple criterion based on the mutual information. This criterion is not sensitive to a pixel permutation, and then it favors the algorithms related to the source PDF. A better criterion taking into account surrounding pixels is needed. This criterion could be therefore the basis of a new algorithm, the one which minimizes its value.

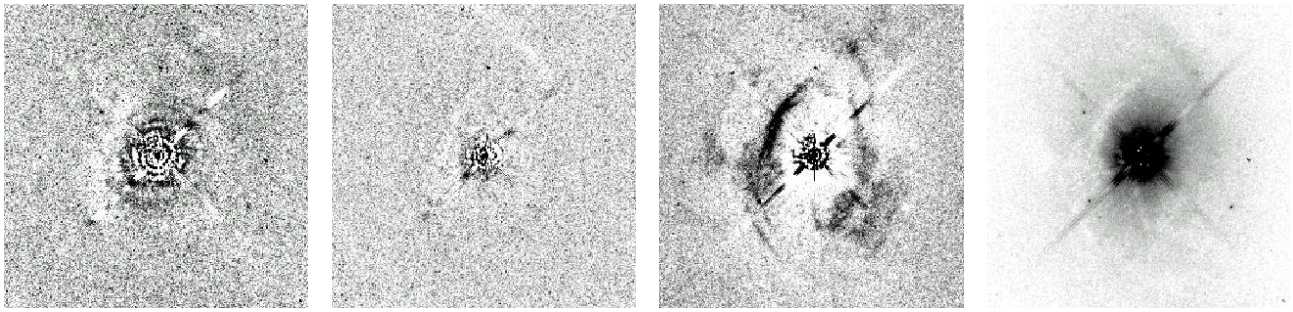

Figure 6: BSS using SOBI4 algorithm with 8 cross correlation matrices.

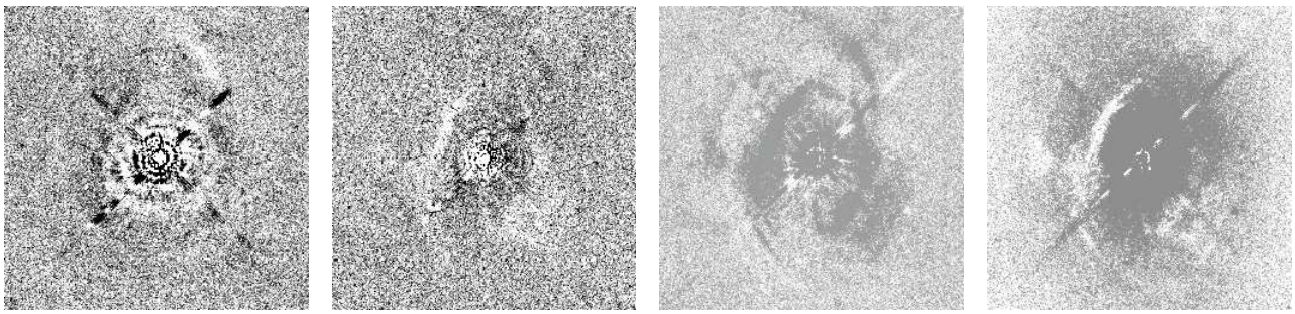

Figure 7: BSS using FastICA algorithm with a deflation algorithm and $g(u) = \tanh(u)$.

5. REFERENCES

[1] A. Bijaoui, V. Doazan, 'Analysis of rapid variations in spectra of α Col by Cross Correlation', *Astron. and Astrophys.*, 70, pp. 285-291, 1979.

[2] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, 'A blind source separation technique using second-ordre statistics,' *IEEE Trans. SP,* 45, pp. 434-444, 1997.

[3] J.F. Cardoso, A. Souloumiac, 'Blind Beamforming for non-gaussian signals', IEE Proceedings-F, 40(6), pp. 362-370, 1993

[4] J.F. Cardoso, A. Souloumiac, 'Jacobi angles for simultaneous diagonalization', *SIAM J. Mat. Anal. Appl.*, 17, pp. 161-164, 1996.

[5] A. Hyvärinen, 'Fast and robust fixed-point algorithms for independant component analysis,' *IEEE Transactions on Neural Networks,* 10(3) pp. 626-634, 1999.

[6] G. Lelièvre, G. Wlérick, J. Sebag, A. Bijaoui, 'Jet optique associé au jet radio de 3C120' *Comptes Rendus Acad. Sciences II*, 318, pp. 905-912, 1994.

[7] F. Murtagh, A. Heck, in *Multivariate data analysis*, Chapter 2, pp. 13-33, Reidel, Dordrecht, 1987.

[8] F. Murtagh, J.L. Starck, A. Bijaoui, 'Image restauration with noise suppression using the Wavelet Transform II,' *Astron. Astrophys. sup. ser.*, 112, pp. 179-189, 1995.

[9] D. Nuzillard, 'Adaptation de SOBI à des données fréquentielles', *GRETSI'99 Vannes*, pp. 745-748, 1999.