Separation of Dependent Sources in Astrophysical Radiation Maps using Second Order Statistics

L. Bedini, S. Bottini, D. Herranz, E.E. Kuruoglu, E. Salerno and A. Tonazzini

Istituto di Scienze e Tecnologie dell’Informazione
CNR - Area della Ricerca di Pisa. Via G. Moruzzi 1, 56124 Pisa (Italy)

Abstract

This paper proposes a semi-blind approach for the separation of sources that are not totally uncorrelated. Up to now, all work on separation of components in astrophysical maps have assumed statistically independent sources. In this work, we attempt to perform a dependent component analysis of astrophysical sources and we propose a semi-blind algorithm which provides separation using second-order statistics only.

1. INTRODUCTION

This paper proposes a semi-blind approach for the separation of sources that are not totally uncorrelated. Up to now, all work on separation of components in astrophysical maps have assumed statistically independent sources. However, it is well known that this is not always true. For example, in sub-mm observations it is known that the galactic foregrounds are not completely independent. In this work, we attempt to perform a dependent component analysis of astrophysical sources and we propose a semi-blind algorithm which provides separation using second-order statistics only. Our prior knowledge about the elements of the mixing matrix are utilised to reduce the required number of equations hence enabling a unique solution by a second-order technique only. Unlike other second-order techniques proposed in the literature, our approach is effective even if some sources are correlated. The approach allows the unknown parameters of the mixing matrix and the source covariance matrices at different shifts to be estimated. Our estimating algorithm has been tested with a database that simulates the maps expected from the instruments that will operate onboard ESAs Planck Surveyor Satellite to measure the CMB anisotropies all over the celestial sphere. We performed separation in several sky patches, featuring different levels of galactic contamination of the CMB, and assuming various noise levels, including the ones derived from the Planck specifications. In all the cases, the CMB reconstruction was satisfactory on all the angular scales considered in the simulation; the average performance of the algorithm and its dispersion were checked and quantified against a large set of noise patterns.

Any radiometric measurement on sky emission results from a superposition of different radiating sources. Separating the individual radiations from the measured signals is thus a common problem in astrophysical data analysis. As an example, in Cosmic Microwave Background (CMB) anisotropy surveys, the cosmological signal is normally combined with radiation of both extragalactic and galactic origin, such as the Sunyaev-Zel'dovich effects from clusters of galaxies, the effect of the individual radiogalaxies, the emission from galactic dust, the galactic synchrotron and free-free emissions. All of these source signals have an interest of their own, and it could be useful to extract all of them from multichannel data, by exploiting their different emission spectra.

Several different component separation techniques have been proposed in the literature. Some authors (Hobson et al. 1998, Bouchet et al. 1999) have tried to achieve the component separation assuming that the physical mixture model is perfectly known. Unfortunately, such an assumption is rather unrealistic and could overconstrain the problem, thus leading to unphysical solutions. Attempts have been made to avoid this shortcoming by introducing criteria to evaluate a posteriori
the closeness to reality of the mixture model and allowing individual sources to be split into separate templates to take spatial parameter variability into account (Jones et al. 1998, Barreiro et al. 2003).

A class of techniques capable of estimating the source signals as well as identifying the mixture model has recently been proposed in astrophysics (Baccigalupi et al. 2000, Maino et al. 2002, Baccigalupi et al. 2002, Delabrouille et al. 2002). In digital signal processing, these techniques are referred to as blind source separation (BSS) and rely on statistical assumptions on the source signals. In particular, mutual independence and nongaussianity of the source processes are often required (Hyvärinen and Oja 2000). This totally blind approach, denoted as independent component analysis (ICA), has already given promising results, proving to be a valid alternative to assuming a known data model. On the other hand, most ICA algorithms do not permit to introduce prior information. Since all available information should always be used, semi-blind techniques are being studied to make astrophysical source separation more flexible with respect to the specific knowledge often available in this type of problem (Kuruoğlu et al. 2003).

The first blind technique proposed to solve the separation problem in astrophysics (Baccigalupi et al. 2000) was based on ICA, allowing simultaneous model identification and signal estimation to be performed. In order to get independence, the statistics of all orders should be taken into account, and this can be achieved by explicit cumulant calculation or by suitable nonlinear transformations on the estimated source signals (Comon 1994, Cardoso 1998, Hyvärinen and Oja 2000).

In general, the problem of estimating all the model parameters and source signals cannot be solved by just using second-order statistics, since these are only able to enforce uncorrelation. However, this has been done in special cases, where additional hypotheses on the spatial correlations or, equivalently, on the spectra of the individual signals are assumed (Tong et al. 1991, Belouchrani et al. 1997, Delabrouille et al. 2002).

A fundamental point in all the previous methods is the assumption of mutual independence of the sources. However, this assumption fails in many relevant cases. For example, it is known that the different Galactic foregrounds (dust, synchrotron and free-free emissions) are not totally independent among themselves. This is particularly true in the strongly contaminated areas of the Galactic plane, where most separation component techniques based on the independence assumption fail.

Another possibility of solving the separation problem on the basis of second-order statistics alone is when the relevant constraints are such that the total number of parameters to be estimated can be reduced. This is the case in the problem of separating astrophysical foregrounds from cosmic microwave background, provided that a structure for the source covariance matrix is assumed. Indeed, on the basis of a possible parametrization of the mixing operator and assuming that the correlation structure between the different sources is known, it is possible to reduce the total number of model parameters. We will show that, under these hypotheses, a very fast model learning algorithm can be devised, based on matching the theoretical covariance matrix at several shifts with the one estimated from the available data. Moreover, this strategy allows the strict assumption of uncorrelation between the source signals to be relaxed. Using this approach, we will show that it is possible to learn the parameters of the mixing even when the sources are significantly correlated among themselves, that is, we provide here for the first time a semi-blind dependent component separation technique applied to astrophysics.

This paper is organized as follows. In Section 2, we formalize the problem and introduce the relevant notation. In Section 3, we describe our method, and, in Section 4, we present some experimental results. Some remarks and future directions are reported in the final section.

2. ASTROPHYSICAL SOURCE SEPARATION

Let us consider an experiment that observes the sky at \( M \) different frequencies. As usual, we model the observations as

\[
x(\xi, \eta) = As(\xi, \eta) + n(\xi, \eta),
\]

(1)
which means that the total radiation observed in a certain direction at a certain frequency is given
by the sum of a number \( N \) of signals (processes, or components). Here, \( \xi \) and \( \eta \) are angular
coordinates on the celestial sphere, \( x = \{ x_d, \, d = 1, \ldots, M \} \) is the \( M \)-vector of the observations,
\( d \) being a channel index, \( A \) is an \( M \times N \) matrix whose entries, \( a_{dc} \), are related to the
spectra of the radiation sources and the frequency responses of the measuring instruments on the
different frequency channels, \( c \) being a process index, \( s = \{ s_e, \, e = 1, \ldots, N \} \) is the \( N \)-vector
of the individual source processes and \( n = \{ n_d, \, d = 1, \ldots, M \} \) is the \( M \)-vector of instrumental noise.
Hereafter, matrix \( A \) will be referred to as the mixing matrix. For simplicity we assume that the
effects of the telescope beam on the angular resolution at different measurement channels have
been equalized (Salerno et al. 2000).

The separation problem consists in estimating the source vector \( s \) from the observed vector \( x \).
Several estimation algorithms have been derived assuming a perfect knowledge of the mixing
matrix. As already said, however, this matrix is related both to the instrumental frequency
responses, which are known, and to the emission spectra \( F_{c}(\nu) \), which are normally unknown.
For this reason it is often necessary to estimate both the mixing matrix and the source vector.
For any procedure of this type, the estimation of \( A \) will be referred to as system identification (or
model learning), and the estimation of \( s \) will be referred to as source separation. In this paper,
we will only emphasize aspects related to learning; indeed, once the model has been identified, a
number of standard reconstruction procedures are available to separate the individual sources.

Assuming that the source signals are mutually independent, the \( MN \) unknown coefficients can
easily be estimated by finding a linear mixture that, when applied to the data vector, nullifies the cross-
cumulants of all orders. If, however, some prior information allows us to reduce the number
of unknowns, the identification problem can be solved by only using second-order statistics.
Learning the model on the basis of second-order statistics alone is computationally easier and it is
also advantageous with respect to robustness against noise; indeed, estimation of second-order
statistics is much more immune from erratic data than estimation of higher-order statistics. This
is the case with our approach, which is based on a parametrization of matrix \( A \). This approach,
described in the following section, also works in the presence of correlated sources.

The elements of \( A \) are related to the source spectra and to the instrumental frequency responses
of the detectors. While the latter is usually very well known, the former is in many cases uncertain.
As we will see, in our case we have some knowledge about them. If we assume that the source
spectra are constant within the passbands of the different channels, we can write the elements
of the mixing matrix as \( a_{dc} \propto F_{c}(\nu_d) \), where \( \nu_d \) is the center-frequency of the \( d \)-th channel and \( F_{c} \)
is the frequency spectrum of the \( c \)-th process.

In the microwave range, the dominant radiations are the CMB, the galactic dust, the free-free
emission and the synchrotron (see De Zotti et al., 1999). Another significant signal comes from
the extragalactic radio sources. Here we assume that the latter has been removed from the data
by one of the specific techniques proposed in the literature (Tenorio et al., 1999, Cayón et al.,
2000, Vielva et al., 2001). Regarding the other components, our knowledge about their frequency
spectra is as follows: The emission spectrum of the CMB is perfectly known, being a blackbody
radiation. In terms of antenna temperature, it is:

\[
F_{\text{CMB}}(\nu) = \frac{\nu^2 \exp(\nu)}{[\exp(\nu) - 1]^2},
\]

(2)

where \( \nu \) is the frequency in GHz divided by 36.8. From (2), the column of \( A \) related to the CMB
radiation is thus known up to an inessential scale factor. For the synchrotron radiation, we have

\[
F_{\text{syn}}(\nu) \propto \nu^{-n_s}.
\]

(3)

Thus, the column of \( A \) related to synchrotron only depends on a scale factor and the spectral
index \( n_s \). For the thermal galactic dust, we have

\[
F_{\text{dust}}(\nu) \propto \frac{\nu^{n+1}}{\exp(\nu) - 1},
\]

(4)
where $\nu = h\nu/kT_{\text{dust}}$ and where $h$ is the Planck constant, $k$ is the Boltzmann constant and $T_{\text{dust}}$ is the physical dust temperature. If we assume a uniform temperature value, the frequency law (4), that is, the column of $A$ related to dust emission, only depends on a scale factor and the parameter $m$. In general, if we assume to have a perfectly known source (such as CMB) and $N - 1$ sources with one-parameter spectra, the number of unknowns in the identification problem is $N - 1$ instead of $N M$. For the sake of simplicity, although other foregrounds (such as SZ and free-free) could be taken into account, in our experiments we only considered synchrotron and dust emissions, which are the most significant in the Planck frequency range.

3. SECOND-ORDER IDENTIFICATION ALGORITHM

Let us consider the source and noise signals in (1) as realizations of two stationary vector random processes, whose components are mutually independent. The covariance matrices of these processes are, respectively,

\[
\begin{align*}
C_s(\tau, \psi) &= \langle [s(\xi, \eta) - \mu_s][s(\xi + \tau, \eta + \psi) - \mu_s]^T \rangle, \\
C_n(\tau, \psi) &= \text{diag} [\sigma_1, \ldots, \sigma_n], \delta(\tau, \psi), \quad \sigma_n = \langle n_i^2 \rangle,
\end{align*}
\]

where $\langle \rangle$ denotes expectation under the appropriate joint probability, $\mu_s$ is the mean vector of the $s$ process, the superscript $T$ means transposition, $(\tau, \psi)$ is the shift at which the covariance matrix are calculated and $\delta$ is the two-dimensional Dirac distribution. The structure in eq. (6) assumes that noise is stationary, uncorrelated across detectors (hence the diagonality) and has no significant auto-correlation (so the covariance matrix is zero for non-zero shifts). We will see later that the robustness of our method against noise allows it to work well even under non-stationary noise conditions. As usual, the noise process is assumed signal-independent, white and zero-mean, with known variances. Note that since mutual independence among the sources is not assumed, $C_n$ is not necessarily a diagonal matrix.

By exploiting equation (1), the covariance of the observed data at any shift can be written as:

\[
C_x(\tau, \psi) = AC_s(\tau, \psi)A^T + C_n(\tau, \psi). 
\]

Let us now define the matrix

\[
H(\tau, \psi) = AC_s(\tau, \psi)A^T = C_x(\tau, \psi) - C_n(\tau, \psi). 
\]

As already proved (Belouchrani et al. 1997, Barros and Cichocki 2001), covariance matrices, i.e. second-order statistics, permit blind separation to be achieved when the sources show a spatial structure, namely, when they are spatially correlated. In these cases, the independence requirement of the ICA approach (Comon 1994) is replaced by an equivalent requirement on the spatial structure of the signal. In other words, finding matrices $A$ and $C_s$ is generally not possible from equation (8) at zero shift alone; higher-order statistics or the covariance matrices at several shifts $(\tau, \psi)$ must be taken into account (Belouchrani et al. 1997).

As assumed in the previous section, in this application we are able to reduce the number of unknowns by parameterizing the mixing matrix. For example, let $A$ only depend on $N - 1$ parameters $\alpha_i$, $i = 1, \ldots, N - 1$, and $C_s$ on its $N \times N$ elements. By introducing a large enough number $L$ of different shifts $(\tau_k, \psi_k)$, $k = 1, \ldots, L$ we will increase the number of nonlinear equations that can be used to solve the problem.

Actually, we do not possess matrix $H$ as defined by (7) and (8), but we are able to get an estimate of it from the known matrix $C_n$ and the following estimate of $C_x(\tau, \psi)$:

\[
C_x(\tau, \psi) = \frac{1}{N_p} \sum_{\xi, \eta} [x(\xi, \eta) - \mu_x][x(\xi + \tau, \eta + \psi) - \mu_x]^T,
\]

where the summation is extended over all the $N_p$ available data samples, and $\mu_x$ is the mean vector of the data sequence. Let

\[
\hat{H} = C_x - C_n.
\]
be our estimate of matrix $\mathbf{H}$ (the indexes $(\tau, \psi)$ have been dropped for simplicity). Matrices $\mathbf{A}$ and $\mathbf{C}_s(\tau, \psi)$ can be estimated from

$$\begin{equation}
\arg \min \left[ \sum_k \| \mathbf{A}(\alpha_1, \ldots, \alpha_{N-1}) \mathbf{C}_s(\tau_k, \psi_k) \mathbf{A}^T(\alpha_1, \ldots, \alpha_{N-1}) - \mathbf{H}(\tau_k, \psi_k) \| \right],
\end{equation}$$

where the minimization is performed over all the possible values of the mixing matrix parameters $\alpha_1, \ldots, \alpha_{N-1}$ and all the possible elements of the $L$ covariance matrices $\mathbf{C}_s(\tau_k, \psi_k)$. An appropriate matrix norm should be selected. Our present strategy to find the minimizer is to alternate a componentwise minimization in $\alpha_k$ with fixed $\mathbf{C}_s$, and the evaluation of $\mathbf{C}_s$, whose elements can be calculated exactly once $\mathbf{A}$ is fixed. A more accurate minimization strategy is now being studied.

Once obtained the elements of $\mathbf{A}$ there are different techniques that can be used to recover the sources (actually attaining separation), their probability densities and any other interesting quantities such as the power spectrum of the CMB, etc. The estimated covariance matrices $\mathbf{C}_s$ at different shifts could be used as well to improve the separation. We will explore this interesting possibility in a future work. For the moment, here we will only focus in the learning step of the source separation.

### 4. EXPERIMENTAL RESULTS

We tested our technique against a simulated data set that is a simplified version of the one expected from the Planck surveyor satellite (Mandolesi et al., 1998, Puget et al., 1998, see also the Planck homepage\(^1\)). The simulations have been made available as a database to the Planck community by the Planck Technical Working Group 2.1. (diffuse component separation). For the sake of simplicity we will consider only three different sources ($N = 3$): the CMB anisotropy, the galactic synchrotron and thermal dust emissions over the four measurement channels ($M = 4$) centered at 30 GHz, 44 GHz, 70 GHz and 100 GHz. The test data maps have been generated by extracting several sky patches at different galactic coordinates from the simulated database, scaling them exactly according to formulas (2)-(4) and generating the mixtures for the channels chosen. We added realizations of noise; to make the simulations more realistic we used a noise pattern such as the like expected for Planck, that is, signal independent noise that is pointwise Gaussian, but whose rms is not stationary across the image. Several noise levels were tested, from the expected Planck levels up to ten times more noise. Figure 1 shows the 100 GHz templates of the three astronomical sources and the noise rms pattern for one of the sky patches we considered. In this section, we report a specific example and describe the results of extensive trials on different sky patches.

For this example we choose a patch located on a strongly contaminated area of the Galactic plane (see figure 1). The instrumental noise levels were comparable to the ones expected for Planck (roughly speaking, this corresponds to a rms of the noise that is $\sim 10\%$ of the rms of the CMB at 100 GHz, see the Planck web site for more detailed specifications). The true mixing matrix, $\mathbf{A}_o$, was derived from equations (2)-(4), with spectral indexes $n_s = 2.9$ and $m = 1.8$ (see for example Banday and Wolfendale, 1991, and Giardino et al. 2002):

$$\mathbf{A}_o = \begin{pmatrix}
1 & 1 & 1 \\
1.135 & 2.813 & 0.548 \\
1.224 & 10.814 & 0.246 \\
1.257 & 32.836 & 0.126
\end{pmatrix}.$$  

We tested our semi-blind method using a $5 \times 5$ grid of shifts, with each step in $\tau$ and $\psi$ corresponding to a distance of 5 pixels. This number of shifts is large enough to allow us to solve the problem and obtain an estimation of the spectral indexes, $\tilde{n}_s = 2.899$ and $\tilde{m} = 1.798$.

\(^1\)http://astro.estec.esa.nl/SA-general/Projects/Planck/
This leads to an estimation of the mixing matrix:

$$\hat{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1.1353 & 2.8124 & 0.5489 \\ 1.2241 & 10.8057 & 0.2468 \\ 1.2570 & 32.7987 & 0.1263 \end{pmatrix}.$$  \hspace{1cm} (13)

We obtain as well the covariance matrices of the sources at the $5 \times 5$ considered shifts. Figure 2 shows the agreement between the true value of the elements of the covariance matrices (blue lines) and the estimated values (red lines). The agreement both in $A$ and $C_s$ is very good even considering that the area is located in the Galactic plane, the correlation between dust and synchrotron is very strong and the noise is not uniform across the image.

To show the robustness of the method against noise, we performed a second set of simulations introducing ten times more noise. In that case, the spectral indexes $\hat{n}_s = 2.906$, $\hat{n}_m = 1.787$ and the mixing matrix

$$\hat{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1.1353 & 2.8196 & 0.5511 \\ 1.2241 & 10.8699 & 0.2491 \\ 1.2570 & 33.0852 & 0.1281 \end{pmatrix}.$$  \hspace{1cm} (14)

are still very well estimated. The estimation of the covariance matrices is comparable to the previous case as well.

We repeated this analysis for several patches and different noise levels. In all the patches taken into account the average parameters estimated are the same as the ones reported above. It is to remark that model learning performs better in patches taken at low galactic latitudes. At high galactic latitudes the only dominant signal is the CMB, and the foregrounds are often well below the noise signal, which makes it difficult for the method to learn the parameters of the foregrounds.
Separation of Dependent Sources

FIGURE 2: True (blue) and estimated (red) value of the elements of the covariance matrices in our test example. Each panel represents an element of the covariance matrix; panels are arranged in the same way as the matrix elements: panels in the diagonal represent autocorrelations (top-left panel stands for CMB, central panel for synchrotron, bottom-right panel for dust), off-diagonal panels show crosscorrelations among different components. Note that off-diagonal terms corresponding to the crosscorrelation between CMB and the Galactic foregrounds show very low values, whereas there is a significant correlation between dust and synchrotron. Inside each panel, the $5 \times 5$ points corresponding to the different shifts have been arranged in one dimension to facilitate visualization.

5. REMARKS

Obviously, the semi-blind identification technique described here cannot be seen as a general approach to separation. However, it relies on a parametric knowledge of the emission spectra, a fairly common assumption in astrophysical data analysis. Our approach permits a very fast and robust model identification, thus enabling an accurate source estimation procedure to be implemented. We also envisage a method to estimate the source covariation matrices at different shifts, which, in their turn, can help the separation task.

Source separation by our method has been particularly interesting with data from low galactic latitudes, where the foreground level is often higher than the CMB signal. Moreover, at these latitudes the statistical dependence between the different Galactic foregrounds is very strong. Note that many separation strategies, both blind and non-blind, have failed their goal in this region of the celestial sphere. As an example, WMAP data analysis (see Bennett et al., 2003) was often performed by using pixel intensity masks that exclude the brightest sky portion from being considered.

This is the first method for semi-blind separation of correlated sources ever proposed for astrophysical purposes. Recently, some methods for a complete blind separation of correlated sources have been proposed in the literature (Barros, 2000). Their effectiveness in astrophysical map separation has not been proved yet. Moreover, they have a high computational complexity.

A problem that is still open with the expected Planck data is the different resolution of the data maps in some of the measurement channels. The identification part of our method can work with maps whose resolution has been degraded in order to be the same in all the channels. The result should be an estimate of the mixing matrix, which can be used in any non-blind separation approach with channel-dependent resolution, such as maximum entropy. However,
the possible asymmetry of the telescope beam patterns should be taken into account in verifying this possibility.

ACKNOWLEDGMENTS

This work has been partially supported by the Italian Space Agency, under contract ASI/CNR 1R/073/01. DH is supported by the European Community’s Human Potential Program under contract HPRN-CT-2000-00124 CMBNET. The authors adopted the simulated sky templates provided by the Planck Technical Working Group 2.1 (diffuse component separation). In particular, the authors are grateful to Martin Reinecke (MPA), Vlad Stolyarov (Cambridge), Andrea Moneti, Simon Prunet and François Bouchet (IAP) for setting up and distributing the database. Extensive use of the HEALPix scheme (Hierarchical, Equal Area and Iso-Latitude Pixelisation of the sphere\(^2\)), by Krysztof M. Górski et al, has been made throughout this work.

REFERENCES

[22] Puget J. L. et al., 1998, proposal submitted to ESA for the Planck High Frequency Instrument

\(^2\)http://www.eso.org/science/healpix