Source separation techniques applied to astrophysical maps

E. Salerno[1], A. Tonazzini[1], E. E. Kuruoğlu[1], L. Bedini[1], D. Herranz[1], C. Baccigalupi [2]

[1]ISTI-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy.
[2]SISSA-ISAS, Via Beirut 4, 34014 Trieste, Italy.

Abstract. This paper summarises our research on the separation of astrophysical source maps from multichannel observations, utilising techniques ranging from fully blind source separation to Bayesian estimation. Each observed map is a mix of various source processes. Separating the individual sources from a set of observed maps is of great importance to astrophysicists. We first tested classical fully blind methods and then developed our approach by adopting generic source models and prior information about the mixing operator. We also exploited a Bayesian formulation to incorporate further prior information into the problem. Our test data sets simulate the ones expected by the forthcoming ESA's mission *Planck Surveyor Satellite* mission.

1 Introduction

To recover the different components from sets of observed maps is an important problem in astrophysics. A radiometric map observed at any frequency band is a combination of emissions received from different sources, whose radiative properties, which affect the coefficients of the combination, are only partially known. Among the various components, the cosmic microwave background (CMB) is of utmost importance since its anisotropies give information about cosmological parameters which would allow a choice among competing cosmological theories to be made. The other components, or *foregrounds*, can be of galactic or extragalactic origin, and each of them has its own interest. Thus, rather than just filtering out the foregrounds, our aim is to extract each individual source. Finding efficient separation methods is an important issue, since an increasingly vast amount of radiometric observations is being made available by current or planned observational missions. Any totally blind source separation (BSS) problem cannot have a unique solution from the observed data alone, since both the coefficients and the sources are to be estimated, and this leads to an unsolvable ambiguity. To eliminate it, one should rely on known source properties. One such approach is independent component analysis (ICA) [6], which assumes mutual independence between the source signals. Even though in principle the astrophysical sources are not mutually independent, we attempted ICA techniques to assess their tolerance to an imperfect data model. We also tested an independent factor analysis method (IFA) [1], which allowed us to treat space-varying noise. In this paper, we summarize and comment our experiments in astrophysical map separation, with particular reference to the data that will be made available by the ESA's *Planck Surveyor Satellite* [11].

2 Data and source models

A common assumption in astrophysical source separation [2] is that each radiation process has a spatial pattern $s_j(\xi, \eta)$ that is independent of its frequency spectrum $F_j(\nu)$, where ξ and η are angular coordinates on the celestial sphere, and ν is frequency. The total radiation observed in a certain direction at a certain frequency is given by the sum of a number N of signals of the type above. Ignoring the effects of the different telescope beams at different frequencies, the observed signal at M distinct channels can be modelled as

$$\mathbf{x}(\xi, \eta) = \mathbf{A}\mathbf{s}(\xi, \eta) + \mathbf{n}(\xi, \eta) \tag{1}$$

where $\mathbf{x}=\{x_i,\ i = 1, \ldots, M\}$ is the M-vector of the observations, i being a channel index, $\mathbf{A}$ is an $M \times N$ matrix whose entries, A_{ij}, are related to the spectra $F_j(\nu)$ of the radiation sources and the frequency responses of the measuring instruments on the different channels, $\mathbf{s} = \{s_j,\ j = 1, \ldots, N\}$ is the N-vector of the individual source processes and $\mathbf{n}=\{n_i,\ i = 1, \ldots, M\}$ is the M-vector of instrumental noise. This noise is normally Gaussian and space-varying. A strictly blind source separation approach assumes $\mathbf{A}$ totally unknown and aims at obtaining A_{ij} and s_j from the observations x_i alone. In our application, however, each column of $\mathbf{A}$ has a known relationship to one of the source spectra $F_j(\nu)$, which either depends on a single unknown spectral index or is completely known (this is the case for CMB and the Sunyaev-Zeldovich radiation from clusters of galaxies). Each element of $\mathbf{A}$ has thus the form $A_{ij} = c_j g_j(\nu_i; \beta_j)$, where c_j is independent of frequency, g_j is a known function of frequency and of a possibly unknown spectral index β_j. ν_i is the center frequency of the i-th channel. Since our problem can be solved up to a scaling ambiguity [6], we can assume a modified matrix whose generic element is

$$\tilde{A}_{ij} = \frac{A_{ij}}{A_{1j}} = \frac{g_j(\nu_i; \beta_j)}{g_j(\nu_1; \beta_j)} \tag{2}$$

The data we used to test all the methods revised below have been either simulated or obtained by extrapolating existing data sets to the frequency range and angular resolution expected for the forthcoming *Planck* data.

3 Fully blind approaches and independent component analysis

As the first step of our exploration, we assumed a noiseless model and adopted the fully blind ICA approach, proposing the first BSS technique to solve the separation problem in astrophysics [2], although in an highly idealised setting. The separation device was a feed-forward neural network that updates the entries of an $N \times M$ separation matrix $\mathbf{W}$ at each received sample of the observed signal. The learning algorithm was a uniform gradient search aimed to minimize the Kullback-Leibler divergence between the probability density $p_{\mathbf{u}}(\mathbf{u})$ of the output vector $\mathbf{u}(\xi, \eta) = \mathbf{W}\mathbf{x}(\xi, \eta)$ and a function $q(\mathbf{u})$, that should represent the factorized joint probability density of the true sources. Since the true source densities are usually unknown, function q is simply chosen according to the supposed sub-Gaussianity (i.e., negative kurtosis) or super-Gaussianity (positive kurtosis) of the sources. The results obtained by this method were quite promising in terms of source estimation accuracy. The accuracies in the estimation of the columns of $\mathbf{A}$ were comparable to the ones obtained for the related sources. The robustness against noise of this technique is not high.

In a successive work [10], we included noise in our model and investigated the performance of the noisy FastICA algorithm [5]. The telescope beam was still assumed frequency independent, and the Gaussian instrumental noise was assumed space-invariant, but at the mean nominal levels for Planck. In terms of accuracy and computational speed, this approach was a considerable improvement over the neural algorithm described above.

An alternative method to deal with noisy mixtures is *independent factor analysis* (IFA) [1]. IFA employs an analytic source model where the source distributions are mixtures of Gaussians, whose parameters are to be estimated jointly with the mixing matrix. The mixing model also contains the noise covariance matrix, which can be estimated as well. IFA is performed in two steps: in the first one (learning), the mixing matrix, the noise covariance and the source density parameters are estimated via an EM algorithm. In the second step (separation), the sources are estimated by using the densities obtained in the first step. In [7], we developed an extension of the original algorithm that assumes a known and space-dependent noise covariance matrix, and updates the source parameters pixel by pixel to consider the different noise variances. We performed the learning step by simulated annealing instead of expectation-maximization (as was done in [1]). This also made the algorithm flexible enough to introduce prior knowledge about the matrix. Experiments with fixed model parameters gave better results than expectation-maximization, even with low SNRs (e.g. 14 dB), yielding a good convergence to the correct mixing matrix. For low SNRs, some of the mixture-of-Gaussian model parameters were estimated rather poorly. Nevertheless, the maximum-likelihood source estimates were better than the ones obtained by FastICA.

Another approach based on generic source models exploits Markov Random Fields models for describing the local autocorrelation of the individual sources, and relies on Bayesian estimation. Though implemented with no assumption on the mixing matrix, and with fixed hyperparameters for the Markov distribution, this method has already given better results than FastICA, in terms of robustness against, possibly nonstationary, noise [12][9].

To profit from the whole richness of prior information we have available, we have extended our formulation to a full Bayesian one that enabled us to assign priors to the source model parameters, which were instead fixed in IFA. This richness in formulation comes with the price of analytical intractability, which we tackled with numerical techniques, namely, Markov Chain Monte Carlo. This formulation enabled us to obtain posterior densities for the variables of the mixing system and to make inferences about all of the source statistics. In comparison with FastICA, we obtained significantly better results [8], at the price of a much higher computational cost.

We also attempted a Bayesian formulation that extends Kalman filtering to non-Gaussian time series. This approach is called *particle filtering*, and is the first to explicitly address the non-stationarity of the data in a problem of astrophysical map separation.This approach potentially provides the most elaborate formulation to the problem, and our initial experiments have already given very promising results [4].

4 Semi-blind approaches and dependent component analysis

All the blind source separation approaches proposed in the literature assume mutual independence between sources. This assumption is often unphysical. In the case we are examining here, for example, significant cross-correlations between the galactic foregrounds are expected. On the other hand, if we exploit the parametrisation of the mixing matrix described in Section 2, the independence assumption may become unnecessary. Parametrising the mixing matrix also allows us to just use second-order statistics to find a unique solution to both the learning and the separation problems. This is a novelty with respect to the partially or totally blind separation techniques proposed so far in astrophysical data analysis, since parametrisation reduces the number of unknowns and allows some of the correlation coefficients to be estimated as well.

Assuming to know which pairs of sources are correlated to each other, we derived a very fast model learning algorithm, based on matching the theoretical zero-shift data covariance matrix to the corresponding empirical matrix [3]. The unknowns of our problem are the set of parameters specifying matrix $\mathbf{A}$ (see Section 2), plus all the nonzero elements of $\mathbf{C}_s$. From eq. (1), the covariance matrix of the observed data at zero shift is:

$$\mathbf{C}_x(0,0) = \langle [\mathbf{x}(\xi,\eta) - \mu_x]\,[\mathbf{x}(\xi,\eta) - \mu_x]^{\mathrm{T}} \rangle = \mathbf{A}\mathbf{C}_s(0,0)\mathbf{A}^{\mathrm{T}} + \mathbf{C}_n\ . \tag{3}$$

where the angle brackets mean expectation, and $\mathbf{C}_n$ is the noise covariance matrix, which is known and diagonal.

Let us now define the matrix

$$\mathbf{H} = \mathbf{C}_x(0,0) - \mathbf{C}_n = \mathbf{A}\mathbf{C}_s(0,0)\mathbf{A}^{\mathrm{T}}. \tag{4}$$

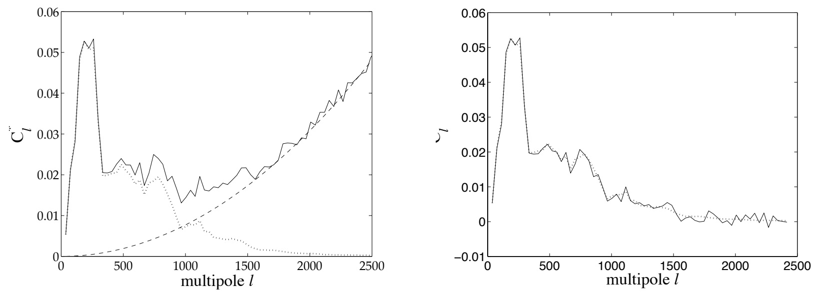

Fig. 1. Left: Real (dotted) and estimated (solid) CMB power spectra, and theoretical noise power spectrum (dashed). Right: Real CMB power spectrum (dotted), and estimated CMB power spectrum corrected for theoretical noise (solid).

An estimate of $\mathbf{H}$ can be derived from matrix $\mathbf{C}_n$ and the sample average of matrices $[\mathbf{x}-\mu_x]\,[\mathbf{x}-\mu_x]^{\mathrm{T}}$ (see eq. 3). Matrices $\mathbf{A}$ and $\mathbf{C}_s$ can then be estimated by minimizing the following form over all the unknowns

$$||\mathbf{H}-\mathbf{A}\mathbf{C}_s(0,0)\mathbf{A}^{\mathrm{T}}||_F \tag{5}$$

where subscript F denotes the Frobenius norm. Of course, the existence of a convenient minimum to (5) is determined by the content of information of the available data set and by the total number of unknowns. In our experiments (see Figures 1 and 2), we learned the model efficiently with four observed maps on a subset of the *Planck* channels, and a 4×3 mixing matrix. Our unknowns were two spectral indices specifying $\mathbf{A}$ and four nonzero elements of the source covariance matrix.

The individual source maps can then be roughly recovered by multiplying the data vectors by the generalised inverse of the mixing matrix estimated. The output maps are corrupted by amplified Gaussian noise with known statistics from which, by deconvolution, we can also recover the source densities. The output noise statistics can also be exploited to improve the accuracy in the estimated angular power spectra of the individual sources, which are of great interest to astrophysicists.

This strategy resulted much more immune from erratic data than the ones based on higher-order statistics. By also taking into account nonzero-shift covariance matrices, it will be possible to estimate the source covariance matrices as functions of the shift. This extension of the method is currently being developed.

In Figure 1, we show how knowledge of the output noise statistics can help in obtaining an accurate power spectrum estimation. In this case we have the spherical harmonic CMB power spectrum for multipoles l from 1 to 2000 (the significant range for the *Planck* observations), obtained by the second-order-statistics method described above. The original data were very noisy (the CMB-to-noise ratio was 0 dB at 100 GHz). It can be seen that by simply subtracting the theoretical noise spectrum from the estimated plot, we can obtain a corrected

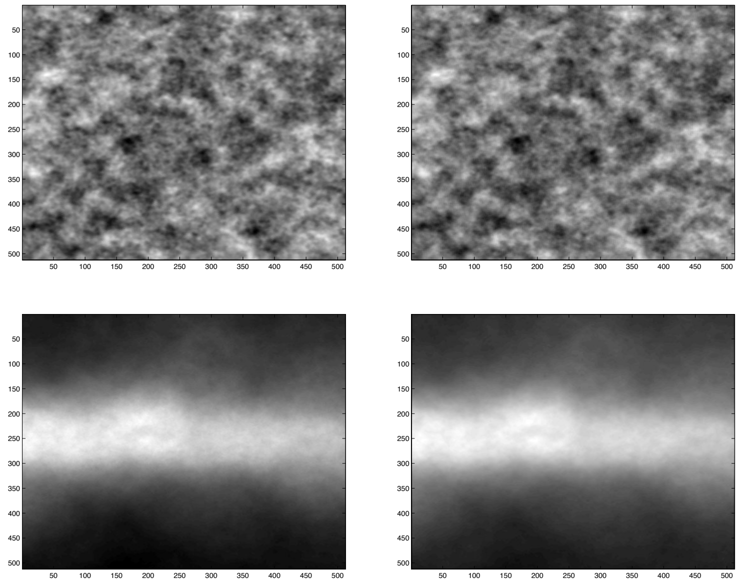

Fig. 2. Source separation based on second-order statistics. Top: original and estimated CMB maps; bottom: original and estimated galactic synchrotron maps.

version of the spectrum. In Figure 2, we see the original and reconstructed CMB and galactic synchrotron maps on a sky patch centered on the galactic plane, where the separation of CMB is normally very difficult. The CMB-to-noise ratio was still 0 dB at 100 GHz. As can be seen, the reconstructed maps are almost indistinguishable from the originals. The results we obtained by this algorithm were the best in terms of robustness against noise, and we assessed them by extensive Monte-Carlo trials [3]. It is also to note that some of the foreground maps in many cases were strongly correlated to each other. This did not impair the learning procedure, from which it was also possible to accurately estimate the cross-correlation coefficients at zero shift.

5 Conclusions

In this paper, we have presented and discussed several blind and semi-blind methods for the separation of components in astrophysical images. Some of them are novel approaches to BSS in imaging, or original extensions of previous methods. In particular, we developed strategies for handling non-stationary noise and for introducing available a priori information into the problem, related to the autocorrelation properties of the individual sources and to the relationships among the mixing coefficients. This brief presentation reflects our path through this challenging problem: starting from the methods based on the pure ICA paradigm, i.e., mutually independent sources and fully blind estimation, we have now achieved significant results in dealing with auto and cross-correlated sources, stationary and non-stationary noise, and in exploiting efficiently the a priori knowledge coming from the physics of the problem.

6 Acknowledgements

D. Herranz is supported by the EC Human Potential Programme under contract HPRN-CT-2000-00124 CMBNET. We are indebted to our colleagues from the *Planck* technical working group 2.1 (diffuse components separation) for fruitful discussions and exchange of results and data. Some of the data processing needed for this work has exploited the HEALPix spherical pixelization scheme (http://www.eso.org/science/healpix), by K. M. Górski et al.

References

1. Attias, H., 1999, Independent factor analysis, Neural Computation, 11, 803.
2. Baccigalupi, C. et al., 2000, Neural networks and the separation of cosmic microwave background and astrophysical signals in sky maps, M.N. Royal Astronomical Society, 318, 769.
3. Bedini, L. et al., 2003, A semi-blind second-order approach for statistical source separation in astrophysical maps, ISTI-CNR, Pisa, Technical Report ISTI-2003-TR-35.
4. Costagli, M., Kuruoğlu, E.E., Ahmed, A., 2003, Bayesian separation of independent components in astrophysical images using particle filters, ISTI-CNR, Pisa, Technical Report, ISTI-2003-TR-54.
5. Hyvärinen, A., Oja, E., 1997, A fast fixed-point algorithm for independent component analysis, Neural Computation, 9, 1483.
6. Hyvärinen, A., Karhunen, J., Oja, E., 2001, Independent Component Analysis, Wiley, New York.
7. Kuruoğlu, E.E. et al., 2003, Source separation in astrophysical maps using independent factor analysis, Neural Networks, 16, 479.
8. Kuruoğlu, E.E., Milani Comparetti, P., 2003, Bayesian source separation of astrophysical images using Markov Chain Monte Carlo, Proc. PHYSTAT, Stanford, 8-11 September 2003.
9. Kuruoğlu, E.E., Tonazzini, A., Bianchi, L., 2004, Source separation in astrophysical images modelled by Markov random fields, submitted to ICIP'04.
10. Maino, D. et al., 2002, All-sky astrophysical component separation with fast independent component analysis (FastICA), M.N. Royal Astronomical Society, 334, 53.
11. http://astro.estec.esa.nl/planck/
12. Tonazzini, A. et al., 2003, Blind separation of auto-correlated images from noisy images using MRF models, Fourth International Symposium on Independent Component Analysis and Blind Source Separation, Nara, Japan, 675.