Extracting Astrophysical Sources from Channel-Dependent Convolutional Mixtures by Correlated Component Analysis in the Frequency Domain

Luigi Bedini and Emanuele Salerno

CNR Istituto di Scienza e Tecnologie dell’Informazione, Via Moruzzi, 1, 56124 Pisa, Italy
{luigi.bedini, emanuele.salerno}@isti.cnr.it

Abstract. A second-order statistical technique (FD-CCA) for semi-blind source separation from multiple-sensor data is presented. It works in the Fourier domain and allows us to both learn the unknown mixing operator and estimate the source cross-spectra before applying the proper source separation step. If applied to small sky patches, our algorithm can be used to extract diffuse astrophysical sources from the mixed maps obtained by radioastronomical surveys, even though their resolution depends on the measurement channel. Unlike the independent component analysis approach, FD-CCA does not need mutual independence between sources, but exploits their spatial autocorrelations. We describe our algorithm, derived from a previous pixel-domain strategy, and present some results from simulated data.

Keywords: Astrophysical imaging, Semi-blind source separation, Dependent component analysis.

1 Introduction

During the last decade, source separation has become an absolute need in cosmological image processing, in view of the growing wealth and accuracy of the observational data sets to be analyzed. The problem consists in extracting the contributions of individual astrophysical sources from the multichannel surveys devoted to the study of the cosmic microwave background (CMB). Since the emission spectra of the astrophysical sources are only known with a coarse approximation, the use of blind source separation techniques has been proposed, which are supposed to both learn the mixing operator and estimate the original source maps.

The first attempts to perform astrophysical source separation by totally blind approaches [1] relied upon the well-known independent component analysis principle, which assures blind separability provided that the mixed sources are mutually independent and nongaussian. This was just an application of known techniques to a specific problem. Indeed, totally blind techniques are not justified in astrophysical source separation, since a relevant prior information is normally available. On the
other hand, assuming mutual independence between the sources is not a good model to what actually happens, since some of the radiation sources are significantly correlated to each other. Moreover, the beamwidths of the radiometric sensors in the millimeter-wave range are normally dependent on the working frequency. This means that the spatial resolutions of the data maps are different in the different channels.

These observations motivated many studies intended to take as much prior information as possible into account, on one hand, and to abandon the independence assumption, on the other. Among the studies amenable to produce methods for the separation of dependent sources, there are the ones that consider second-order statistics and introduce information on the spatial autocorrelation of the individual source maps [2, 3]. As far as the multiple-resolution problem is concerned, the methods working in the pixel domain have been relying on downgrading the resolution to a common value, unlike the frequency-domain methods, which can treat directly the channel-dependent convolutional kernels affecting the data maps.

Our correlated component analysis (CCA) strategy for astrophysical source separation [4] exploits second-order statistics and the spatial autocorrelation within each individual source map. Moreover, it introduces information about the parametric form of the emission laws of the astrophysical sources, which are normally known with a good approximation from their physical models. In this way, we are able to both reduce the number of unknown mixing parameters, and estimate the cross-correlations between the different sources. This approach has already been applied to estimate the spectral indices of the thermal dust and synchrotron galactic radiations from a simulated data set in the range 30 GHz - 545 GHz, contained in the frequency range covered by the forthcoming ESA mission Planck Surveyor Satellite\(^1\) [5]. In our original formulation, CCA works in the pixel space, and solves the problem of the channel-dependent beamwidth by just downgrading the map resolutions to a common (worst) value. Once the mixing operator (i.e., the spectral indices, see below) has been learned, the source estimation problem can be solved from the original data, by using one of the non-blind separation procedures reported in the literature.

One approach used to perform the learning phase by exploiting the available full-resolution maps is to work in a transformed domain. In the case of an astronomical survey, the transformed domain can be the 2D Fourier space if the processing is performed on sky patches that are small enough to be considered flat; otherwise, a set of spherical basis functions is to be used, as in the spherical harmonic transform. In order to integrate the learning and the source estimation phases without the need of downgrading the resolution, we are developing a Fourier-domain version of CCA (FD-CCA) working on sky patches with at most 15° edges. The results obtained by this version are perfectly comparable to the ones obtained by the pixel-domain CCA, with two important advantages: a frequency-dependent beamwidth is naturally accounted for, with no need to preprocess the data, and the source cross-spectra estimated as a byproduct of the learning procedure can be used to help the estimation of the separated source maps via known Bayesian approaches, such as the multifrequency Wiener filtering or the maximum entropy reconstruction [6, 7, 8].

In this paper, we describe FD-CCA and show some results we obtained with simulated data. In Sect. 2, the algorithm is presented. In Sect. 3, we briefly discuss the

\(^1\) Planck webpage: http://planck.esa.int/
data sets used for the simulated experiments and, in Sect. 4, we show some significant experimental results. The concluding section looks at some future developments in astrophysical source separation.

2 FD-CCA Model Learning

Before describing our algorithm, we briefly formalize the separation problem and introduce the notation used throughout this paper.

In our model, the measured data $x$ at a generic pixel $i$ are generated from the underlying components $s$ through a linear, space-invariant, noisy, convolutional mixture operator described as follows [1]:

$$x(i) = (H^*As)(i) + n(i), \quad (1)$$

where $x$ and $s$ are $N$-dimensional and $M$-dimensional vectors, respectively, with $N \geq M$, $A$ is an unknown $N \times M$ space-invariant matrix, $n$ is the $N$-dimensional, signal-independent, noise vector, the asterisk means convolution, and $H$ is an $N \times N$ diagonal matrix whose entries are known convolutional kernels that model the telescope radiation patterns at the related measurement channels. Each 2D pixel index, $i$, represents a particular bearing in the celestial sphere. From the data in (1), we can evaluate the data covariance matrices at lags $\tau$ (if we assume stationary sources, these will not depend on $i$, see also [4]).

$$C_X(\tau) = E[(H^*As)(i)(H^*As)^T(i + \tau)] + C_N(\tau), \quad (2)$$

where the notation $E[\cdot]$ means expectation. If the data maps are so small to be considered flat, Eq. (2) can be translated in the Fourier domain, thus becoming a relationship among cross-spectra estimated by binning the Fourier transforms of the quantities in (2) over concentric annular domains. We have

$$\tilde{C}_X(l) = \tilde{H}(l)A\tilde{C}_S(l)A^T\tilde{H}^\dagger(l) + \tilde{C}_N(l), \quad (3)$$

where the tilde accent denotes the cross-spectrum matrices, $l$ is the generic spatial frequency bin, and the dagger denotes the adjoint matrix.

We reasonably assume that an estimate of the noise cross-spectrum matrix $\tilde{C}_N(l)$ is available. Moreover, to reduce the number of unknowns, we exploit the fact that matrix $A$ only depends on a few spectral indices that form an unknown vector $p$ [5]. Indeed, $A$ is subject to strong physical constraints; for example, it contains a perfectly known column related to the Planck-law blackbody CMB emission spectrum, and a column related to the galactic dust emission, which only depends on a single unknown spectral index and on the physical dust temperature, often assumed known.

Let us now arrange matrix $\tilde{C}_X(l) - \tilde{C}_N(l)$, row by row, in the lexicographically ordered $N^2$-vector $d(l)$, and the source cross-spectrum matrix $\tilde{C}_S(l)$ in the $M^2$-vector $c(l)$. It can be easily shown that (3) can be rewritten as

$$d(l) = A_K(l)c(l), \quad (4)$$

where $A_K$ is the $N^2 \times M^2$ matrix.
Correlated Component Analysis in the Frequency Domain

\[ A_K(l) = [\tilde{H}(l)A] \otimes [\tilde{H}(l)A], \]  

(5)

and the symbol “\( \otimes \)” denotes the Kronecker product. In this way, we have an equation where, for each \( l \), the left-hand side can be estimated from the measured data and the known noise power spectrum, and the right-hand side depends on the unknown spectral indices through matrix \( A_K \) and on the unknown elements of matrix \( \tilde{C}_S(l) \) through vector \( c \). The prior information we introduced is in the parametric structure of matrix \( A \). The source cross-spectrum \( \tilde{C}_S(l) \) is a symmetric matrix, and contains \( M(M+1)/2 \) distinct elements for each \( l \). For any pair of assumedly uncorrelated sources, one of these elements can be kept at zero for all \( l \), thus reducing the number of unknowns. The unknowns \( c(l) \) and \( p \) can be estimated from (4) by solving the following optimization problem:

\[ \hat{c}(l), \hat{p} = \arg \min_{c(l), p} \Phi[c(l), p], \]  

(6)

with

\[ \Phi[c(l), p] = \sum_{l=1}^{l_{\text{max}}}[d(l) - A_K(l, p)c(l)]^T \cdot [d(l) - A_K(l, p)c(l)] + \lambda \Gamma[c(l)]. \]  

(7)

Since the reconstruction of the power spectra is an ill-posed problem, we included the regularization term \( \Gamma \) in the objective functional \( \Phi \), with the usual regularization parameter \( \lambda \). As examples, we can give the two following forms for \( \Gamma \), enforcing smoothness and low total energy, respectively:

\[ \Gamma[c(l)] = \sum_{l=1}^{l_{\text{max}}-1}[c(l+1) - c(l)]^T \cdot [c(l+1) - c(l)], \]  

(8)

\[ \Gamma[c(l)] = \sum_{l=1}^{l_{\text{max}}}c^T(l) \cdot c(l). \]  

(9)

To find the minimum in (6), for any set of spectral indices \( p \), we evaluate vector \( c \) by a standard algorithm (e.g. steepest descent), and then update \( p \) by simulated annealing. The stop criterion is established on the outcome of the annealing schedule, namely, the final evaluation of \( c(l) \) is performed when vector \( p \) has reached a stable value. Since \( A \) is uniquely determined by the spectral indices, estimating \( p \) is tantamount to learning the mixing model. The estimated source cross-spectra \( c(l) \) can then be used to help reconstructing the source maps \( s(i) \) by Bayesian techniques. Note that our \( c(l) \) does not suffer from the bias introduced when the spectra are evaluated from the reconstructed maps (see also [2]).

The present separation approach, as formulated in (6) and (7), starts from the same principle proposed in [2]. Relevant differences are the parametrization of matrix \( A \), and the use of a regularization approach to estimate the source cross-spectra. These features should assure a good determination of the problem (namely, a sufficiently
small number of unknowns for a certain data set) and a good stability of the result. Note also that a smoothness requirement as the one described in (8) comes from a perfectly reasonable assumption for the cross-spectra. Another difference is the Euclidean distance between the lexicographically ordered matrices in (7), instead of the divergence between two positive-definite matrices used in [2].

Fig. 1. Simulated source emission maps (thermodynamic temperature) at 100 GHz, linear grayscales (black=minimum, white=maximum), angular resolution 1.7’. Left: CMB anisotropies, (–0.33 mK, 0.33 mK). Center: galactic synchrotron, (0 mK, 0.2 mK). Right: galactic dust thermal emission, (0 mK, 23 mK).

Fig. 2. Simulated data maps obtained by mixing the sources in Fig. 1 at different frequencies, linear grayscales. Left: 30 GHz map, angular resolution 33’. Center: 100 GHz map, angular resolution 9.2’. Right: 857 GHz map, angular resolution 5’. SNR = 10 dB.

3 Simulated Data

Whereas the pixel-domain CCA is now being used to analyze real radioastronomical data, we are interested in evaluating the performance of FD-CCA as a function of perfectly controllable variables, such as noise, beamwidths, number of relevant sources, individual source emission spectra. For this reason, we are using simulated data. In particular, our data have been prepared by the Planck mission working group on diffuse component separation. We used source maps taken from different sky regions, and their emission spectra, to build simulated data maps at the nine channels expected from the Planck mission, whose center frequencies are 30 GHz, 44 GHz, 70 GHz, 100 GHz, 143 GHz, 217 GHz, 353 GHz, 545 GHz, 857 GHz. In Fig. 1, as an example, we show the maps of the CMB anisotropies, the galactic synchrotron
emission, and the galactic dust thermal emission at 100 GHz, taken from a $15^\circ \times 15^\circ$ sky patch located across the galactic plane, where the foreground emissions are particularly strong. These maps have been mixed by a known operator, then convolved by channel-specific telescope beams and added with noise. In real world, the noise in each channel depends on the sensitivity of the related instrument, and is normally space-variant, because of the uneven sky coverage. In Fig. 2, we show the resulting simulated data maps at three different channels for uniform noise with 10 dB SNR$^2$. The angular resolutions are sensibly worse at low-frequency channels.

![Fig. 2.](image)

**Fig. 3.** Comparison between the reconstructed sources and the originals in Fig. 1. Left: CMB scatterplot. Center: synchrotron scatterplot. Right: dust scatterplot.

## 4 Results Overview

We assessed the performances of FD-CCA by using both the regularization functions (8) and (9), and with different levels of stationary or nonstationary noise. The results found are comparable to the ones obtained by the pixel-domain CCA applied to limited sky patches [4, 5]. At present, we have a set of preliminary results, from which we cannot yet find definite answers on possible resolution improvements obtained by avoiding preprocessing operations on the data maps. Conversely, the advantage of estimating the cross-spectra $c(l)$, mentioned in Sect. 2, has been fully confirmed, since the Wiener-filter source reconstruction from the learned model parameters has given very good results. Indeed [6], the multichannel Wiener reconstruction matrix is

$$W(\omega) = \tilde{C}_S(\omega)A^T \tilde{H}^\dagger(\omega) \left[ \tilde{H}(\omega)A\tilde{C}_S(\omega)A^T \tilde{H}^\dagger(\omega) + \tilde{C}_N(\omega) \right]^{-1},$$

and, obviously, we can expect a good result inasmuch as accurate estimates for $\tilde{C}_S(\omega)$ and $\tilde{C}_N(\omega)$ are available. As mentioned above, our spectrum estimates were made over annular frequency bins. To obtain the spectra as functions of the 2D frequency $\omega$, as required by (10), we assumed that these are circularly symmetric.

For realistic values of the SNR, the reconstructed sources are visually undistinguishable from the originals. This also happens with the data in Fig. 2. To appreciate the quality of our reconstructions, in Fig. 3, we report the scatterplots comparing the estimated and the original sources pixel by pixel.

---

$^2$ Although we are showing results with space-invariant noise, this SNR is well within the actual Planck specifications (details in the Planck webpage).
This result shows that the reconstruction has been very good, especially for CMB. In Fig. 4, we show the CMB power spectrum estimated by FD-CCA, compared to both the original and the one evaluated from the Wiener-filter reconstructed map.

![CMB Power Spectrum](image)

**Fig. 4.** Solid lines: estimated spectra. Dotted lines: spectrum computed from the original in Fig. 1, left panel. Left: CMB power spectrum estimated by FD-CCA from the data partially shown in Fig. 2. Right: spectrum from the Wiener-reconstructed CMB map, reported here to validate our reconstruction.

## 5 Conclusion

We are developing a Fourier-domain version of the already proposed CCA technique to overcome some of the drawbacks that hamper its practical application. Besides the mentioned difficulties of a channel-dependent angular resolution, many other peculiarities characterize the problem of source separation as applied to astrophysical data analysis. Some of them have been treated in different ways during the last few years. Others pose problems that are not completely solved yet. Among them, there are the spatial variability of the noise variance and of the mixing operator $A$. The former is due to the particular sensor scanning strategy that is chosen; the latter depends on the nonuniformity of the physical features of some emission sources, for example, the temperature of the galactic dust grains [8]. One approach to face these problems can be to partition the celestial sphere into a number of limited patches within which uniform features are assumed, and then to mosaicize the separated results to obtain the whole sky again. FD-CCA is suited to follow this approach, although use of the Fourier transform is only allowed when the patches are so small to be considered flat. With larger patches, other basis functions should be used, such as the spherical harmonic functions. Another strategy that could cope with space-variability by using FD-CCA-based algorithms is to adopt a wavelet basis on the whole sphere. In principle, owing to the known properties of wavelets, this should enable us to maintain the global features of the source maps while treating properly the spatial variability [3]. Other strategies to solve this problem rely on processing methods that do not need to compute statistics from the data sample, and are based on Monte Carlo Markov Chain techniques [9, 10]. At present, these techniques cannot be fully exploited because of their computational complexity.
Acknowledgments. Partial support from the Italian Space Agency (I/R/065/04) and the EU Network of Excellence MUSCLE (FP6-507752) is acknowledged. The authors are indebted to the Planck working group on diffuse component separation, especially to Andrea Farusi and Diego Herranz, for preparing the simulated test maps.

References