Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging.

Jean-François Cardoso, CNRS / ENST - TSI, Paris France
Hichem Snoussi, L2S - Supelec, Gif-sur-Yvette, France
Jacques Delabrouille, Guillaume Patanchon PCC - Collège de France, Paris, France

ABSTRACT

We present a new source separation method which maximizes the likelihood of a model of *noisy* mixtures of stationary, possibly Gaussian, independent components. The method has been devised to address an astronomical imaging problem. It works in the spectral domain where, thanks to two simple approximations, the likelihood assumes a simple form which is easy to handle (low dimensional sufficient statistics) and to maximize (via the EM algorithm).

1 SOURCE SEPARATION for ASTRONOMY

1.1 Astronomical components

Source separation consists in recovering components from a set of observed mixtures. Component separation is a topic of major interest to the Planck space mission, to be launched in 2007 by ESA to map the cosmic microwave background (CMB). The blackbody temperature of this radiation as a function of direction on the sky will be measured in $m = 10$ different frequency channels, corresponding to wavelengths ranging from $\lambda = 350\,\mu$microns to $\lambda = 1\,$cm. In each channel, the temperature map will show not only the CMB contribution but also contributions from other sources called *foregrounds*, among which Galactic dust emission, emission from very remote (and hence quasi point-like) galaxy clusters, and several others. It is expected that (after some heavy pre-processing), the map built from the i-th channel can be accurately modeled as $y_i(\vec{r}) = \sum_{j=1}^{n} a_{ij} s_j(\vec{r}) + n_i(\vec{r})$ where $s_j(\vec{r})$ is the spatial pattern for the j-th component and $n_i(\vec{r})$ is an additive detector noise. In other words, cosmologists expect to observe a noisy instantaneous (*i.e.* non convolutive) mixture of essentially independent components (independence being the consequence of the physically distinct origins of the various components). Even though recovering as cleanly as possible the CMB component is the primary goal of the mission, astrophysicists are also interested in the *other* components, in particular for collecting data regarding the morphology and physical properties of Galactic foregrounds (dust...) and the distribution of galaxy clusters.

This paper deals with *blind* component separation. Blindness means recovering the components without knowing in advance the coefficients of the mixture. In practice, this may be achieved by resorting to the strong but often plausible assumption of mutual statistical independence between the components. The motivation for a blind approach is obvious: even though some coefficients of the mixture may be known in advance with good accuracy (in particular those related to the CMB), some other components are less well known or predictable. It is thus very tempting to run blind algorithms which do not require *a priori* information about the mixture coefficients.

1.2 Blind separation methods

Several attempts at blind component separation for CMB imaging have already been reported. The first proposal, due to Baccigalupi *et al.* use a non Gaussian noise-free i.i.d. model for the components[1], hence following the 'standard' path to source separation. One problem with this approach is that the most important component, namely the CMB itself, seems to closely follow a Gaussian distribution. It is well known that, in i.i.d. models, it is possible to accommodate at most one Gaussian component. It does not seem to be a good idea, however, to use a non Gaussian model when the main component itself has a Gaussian distribution.

Another reason why the i.i.d. modeling (which is implicit in 'standard' ICA) probably is not appropriate to our application: most of the components are very much dominated by the low-frequency part of their spectra. Thus sample averages taken through the data set tend not to converge very quickly to their expected values. This may explain why Fourier methods, presented below, seem to perform better.

Thus, rather than exploiting the non Gaussianity of (all but one of) the components, one may think of exploiting their spectral diversity. A very sensible approach is proposed by Pham: using the Whittle approximation of the likelihood, he shows that blind separation can be achieved by jointly diagonalizing spectral covariance matrices computed over several frequency bands [3]. This conclusion however is reached only in

the case of noise-free models. Therefore, it is not appropriate for CMB imaging where a very significant amount of noise is expected.

In this paper, we follow Pham's approach but we take additive noise into account, leading to a likelihood function which is no longer a joint diagonality criterion, thus requiring some new algorithmics. We present below the form taken by the EM algorithm when applied to a set of spectral covariance matrices. This approach leads to an efficient algorithm, much faster than the algorithms obtained via the EM technique in the case of non Gaussian i.i.d. modeling as in [2] or [4].

1.3 A stationary Gaussian model

Our method is obtained by starting from a stationary Gaussian model. For ease of exposition, we assume that the observations are m times series rather than m images (extension to images is straightforward). The $m \times 1$-dimensional observed process $y(t) = [y_1(t); \ldots; y_m(t)]$ is modeled as

$$y(t) = As(t) + n(t) \tag{1}$$

where A is an $m \times n$ matrix with linearly independent columns. The n-dimensional source process $s(t)$ (the components) and the m-dimensional noise process $n(t)$ are modeled as real valued, mutually independent and stationary with spectra $S_s(\nu)$ and $S_n(\nu)$ respectively. The spectrum of the observed process then is

$$S_y(\nu) = AS_s(\nu)A^\dagger + S_n(\nu). \tag{2}$$

The $\dagger$ superscript denotes transconjugation even though transposition would be enough almost everywhere in this paper (our method is easily adapted to deal with complex signals/mixtures). The assumption of independence between components implies that $S_s(\nu)$ is a diagonal matrix:

$$[S_s(\nu)]_{ij} = \delta_{ij} P_i(\nu)$$

where $P_i(\nu)$ is the power spectrum of the ith source at frequency ν. For simplicity, we also assume that the observation noise is uncorrelated both in time and across sensors:

$$[S_n(\nu)]_{ij} = \delta_{ij}\sigma_i^2 \tag{3}$$

meaning that the noise spectral density σ_i^2 on the ith detector does not depend on frequency ν. In summary the probability distribution of the process is fully specified by $m \times n$ mixture coefficients, m noise levels and n power spectra.

2 THE OBJECTIVE FUNCTION

Our method boils down to adjusting smoothed versions of the spectral covariance matrices (2) to their empirical estimates. The estimated parameters are those which give the best match, as measured by an objective function. This objective function is introduced in this section. In the following section, we show how it stems from the maximum likelihood principle.

2.1 Spectral averaging.

A key feature of our method is that it uses low dimensional statistics obtained as averages over *spectral domains* in Fourier space. These Fourier domains simply are frequency bands in the 1D case or are two-dimensional domains of the Fourier plane when the method is applied to images.

Consider a partition of the frequency interval $(-\frac{1}{2}, \frac{1}{2})$ into Q domains (bands): $(-\frac{1}{2}, \frac{1}{2}) = \cup_{q=1}^{Q} \mathcal{D}_q$ which are required to be symmetric: $f \in \mathcal{D}_q \Rightarrow -f \in \mathcal{D}_q$. For any function $f(\nu)$ of frequency, denote $\langle f \rangle_q$ its average over the q-th spectral domain when sampled at multiples of $1/T$:

$$\langle f \rangle_q = \frac{1}{w_q} \sum_{\frac{p}{T} \in \mathcal{D}_q} f\left(\frac{p}{T}\right), \qquad q = 1, \ldots, Q \tag{4}$$

where w_q is the number of points in domain $\mathcal{D}_q$.

2.2 Spectral statistics

Denoting $Y(\nu)$ the discrete-time Fourier transform of T samples:

$$Y(\nu) = \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} y(t) \exp(-2i\pi\nu t), \tag{5}$$

the *periodogram* is $\hat{S}_y(\nu) = Y(\nu)Y(\nu)^\dagger$ and its averaged version is

$$\langle \hat{S}_y \rangle_q = \langle Y(\nu)Y(\nu)^\dagger \rangle_q. \tag{6}$$

Note that $Y(-\nu) = Y(\nu)^*$ for real data so that $\langle \hat{S}_y \rangle_q$ actually is a real valued matrix if $\mathcal{D}_q$ is a symmetric domain.

This sample spectral covariance matrix will be our estimate for the corresponding averaged quantity

$$\langle S_y \rangle_q = A \langle S_s \rangle_q A^\dagger + \langle S_n \rangle_q \tag{7}$$

where the equality results from averaging model (2).

A key point is that the structure of the model is not affected by spectral averaging since $\langle S_s \rangle_q$ remains a diagonal matrix after averaging:

$$\langle S_s \rangle_q = \mathrm{diag}\,[\langle P_1 \rangle_q, \ldots, \langle P_n \rangle_q]$$

and, of course, $\langle S_n \rangle_q = S_n = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ still is a constant diagonal matrix.

2.3 Blind identification via spectral matching

Our proposal for blind identification simply is to match the sample spectral covariance matrices $\langle \hat{S}_y \rangle_q$, which depend on the data, to their theoretical values $\langle S_y \rangle_q$, which depend on the unknown parameters. There are $m \times n + Q \times n + m$ of these parameters, [1] collectively referred to as θ:

$$\theta = \left\{ [A_{ij}]_{i=1,j=1}^{i=m,j=n};\ \ [\langle P_j \rangle_q]_{j=1,q=1}^{j=n,q=Q};\ \ [\sigma_i^2]_{i=1}^{i=m} \right\}. \tag{8}$$

[1] There are actually n redundant parameters since a scale factor can be exchanged between each column of A and the corresponding power spectra.

The mismatch between the sample statistics and their expected values is quantified by the average divergence:

$$\phi(\theta) = \sum_{q=1}^{Q} w_q \; D\left(\langle \hat{S}_y \rangle_q, \langle S_y \rangle_q\right) \tag{9}$$

where the positive weight w_q is (as above) proportional to the size of the q-th spectral domain and where $D(\cdot,\cdot)$ is a measure of divergence between two $m \times m$ positive matrices defined as

$$D(R_1, R_2) = \mathrm{tr}\left(R_1 R_2^{-1}\right) - \log\det(R_1 R_2^{-1}) - m \tag{10}$$

This is nothing but the Kullback divergence between two n-dimensional zero-mean Gaussian distributions with positive covariance matrices R_1 and R_2 respectively.

The reason for using the mismatch measure (9) is its connection to maximum likelihood principle (see below). Even though the divergence (9) may, at first sight, seem more difficult to deal with than a simple quadratic distance, it is actually a better choice in at least two respects: first, we expect it to yield efficient parameters estimates because of the asymptotic optimality of maximum likelihood estimation; second, thanks to its connection to the log-likelihood, it lends itself to simple optimization via the EM algorithm (see below).

A last note: since domain averaging does not change the algebraic structure of the spectral covariance matrices (*i.e.* eq. (2) becomes (7)), it does not introduce any bias in the estimation of A.

3 MAXIMUM LIKELIHOOD AND EM

3.1 Whittle approximation

The Whittle approximation is a spectral approximation to the likelihood of stationary processes. It has been introduced for the blind separation of noise-free mixtures by Pham [3]. Simplifying a little bit, this approximation boils down to asserting that the coefficients $Y(\nu)$ of definition (5) taken at frequencies $\nu = p/T$ are uncorrelated, have zero-mean and a covariance matrix equal to $S_y(\nu)$. Simple computations then show that —up to a constant and a scalar factor— the (negative) log-likelihood of the data takes the form (9) under the additional approximation that $S_y(\nu)$ is constant over each spectral domain.

The Whittle approximation is good for Gaussian processes but certainly does not capture all the probability structure, even for large T, for non Gaussian processes. However, it it still provides a principled way of exploiting the spectral structure of the process, leading to the selecting of (9) as an objective function. In addition, it suggests to use the EM algorithm for minimizing (9).

3.2 An EM algorithm in the spectral domain

Using the EM technique for maximizing a likelihood function requires defining latent (unobserved) data. In the case of source separation, there is an obvious choice: take the components as the latent data. This approach was introduced in [2] for a noisy non Gaussian i.i.d. model of source separation and later in [4] for a noisy non Gaussian model in the spectral domain. Both these models lead to heavy computation. In contrast, by i) using only a Gaussian model (the Whittle approximation) and ii) averaging over spectral domains, the EM algorithm becomes very computationally attractive.

Room is lacking for a complete derivation of the EM algorithm in our case but it is not difficult to adapt, for instance, the computations of [2] to our case. Let us only mention why the resulting algorithm is much simpler.

First, when dealing with data structured as $y = As + n$, EM needs to evaluate conditional expectations $E(s|y)$ and $E(ss^\dagger|y)$. Thanks to the Gaussian model, these are readily found to be *linear* functions of y and $yy^\dagger$ respectively:

$$E(s|y) = Wy \tag{11}$$
$$E(ss^\dagger|y) = Wyy^\dagger W^\dagger + C \tag{12}$$

where matrices C and W are defined as

$$C = (A^\dagger R_n^{-1} A + R_s^{-1})^{-1} \tag{13}$$
$$W = (A^\dagger R_n^{-1} A + R_s^{-1})^{-1} A^\dagger R_n^{-1} \tag{14}$$

with covariance matrices $R_s = \mathrm{Cov}(s)$, $R_n = \mathrm{Cov}(n)$.

Second, this linearity is preserved through domain averaging, meaning that the EM algorithm only needs to operate on the sample covariance matrices $\langle \hat{S}_y \rangle_q$. This set of matrices is a sufficient statistic set in our model; it is also all that is needed to run the EM algorithm.

Thus, blind separation of noisy mixtures of stationary sources can be achieved by computing the periodogram, averaging it into a set of sample covariance matrices and maximizing the likelihood by then running the EM algorithm. The algorithm is summarized in pseudo-code (see Alg. 1), but its derivation (which is purely mechanical) is omitted. In this pseudo-code, the $\mathrm{diag}(\cdot)$ operator sets to 0 the off-diagonal elements of its argument. We also include a renormalization step which deals with the scale indetermination inherent to source separation: each column of A is normalized to have unit norm and the corresponding scale factor is applied to the average source spectra.

4 APPLICATION AND COMMENTS

4.1 Separating astrophysical components

Preliminary tests have been carried on simulated observations in six channels at microwave frequencies 100, 143, 217, 353, 545 and 857 GHz, over sky patches of size $12.5^\circ \times 12.5^\circ$ sampled over a 300×300 pixel grid. Mixtures include three astrophysical components (CMB, Galactic dust, and emission from galaxy clusters) and white noise.

Since isotropic observations are expected, we choose spectral domains which are not only symmetric but also

Start with sample covariance matrices $\langle \hat{S}_y \rangle_q$, and initial guesses for A, S_n and $\langle S_s \rangle_q$.
repeat
  {E-step ———— Compute conditional statistics}
  for $q = 1$ to Q **do**
    $C_q = (A^\dagger S_n^{-1} A + \langle S_s \rangle_q^{-1})^{-1}$
    $R_{yy}(q) = \langle \hat{S}_y \rangle_q$
    $R_{ys}(q) = \langle \hat{S}_y \rangle_q S_n^{-1} A \; C_q$
    $R_{ss}(q) = C_q \; A^\dagger S_n^{-1} \langle \hat{S}_y \rangle_q S_n^{-1} A \; C_q + C_q$
  end for
  $R_{ss} = \frac{1}{T} \sum_{q=1}^{Q} \; w_q R_{ss}(q)$
  $R_{ys} = \frac{1}{T} \sum_{q=1}^{Q} \; w_q R_{ys}(q)$
  $R_{yy} = \frac{1}{T} \sum_{q=1}^{Q} \; w_q R_{yy}(q)$
  {M-step ———— Update the parameters}
  $A = R_{ys} R_{ss}^{-1}$
  $S_n = \mathrm{diag}\left(R_{yy} - R_{ys} R_{ss}^{-1} R_{ys}^\dagger\right)$
  $\langle S_s \rangle_q = \mathrm{diag}\left(R_{ss}(q)\right)$ for $1 \le q \le Q$.
  Renormalize A and the $\langle S_s \rangle_q$
until convergence

Alg. 1: Gaussian EM algorithm over spectral domains

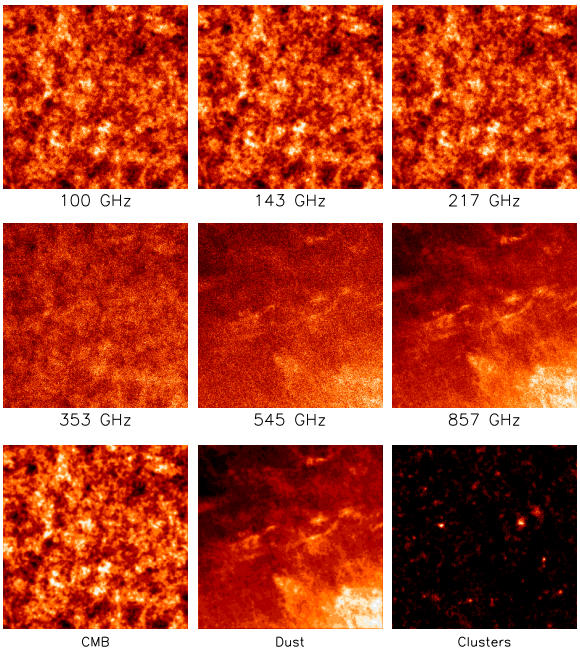

Figure 1: Top two rows: maps at the detectors. Bottom row: components extracted with the Wiener filter based on the estimated parameters.

rotation invariant; in other words: spectral rings. The sample spectral covariance are computed over 15 such rings equally spaced over the whole band. Data reduction thus is by a factor of $(6 \times 300 \times 300)/(15 \times 6 \times 6) = 1000$. The EM algorithm converges in a few tens of iterations amounting to a few seconds on a 1 GHZ machine when coded in octave (a free clone of Matlab: `http://www.octave.org`).

Room is lacking for a detailed description of our experiments which will be reported elsewhere. See figure 4.1 for an illustration with a typical (and significant) level of noise. A notable feature here is the separation of the galaxy clusters. The SNR on the CMB component is also much improved at high frequency even though this cannot be assessed from the picture. See the companion paper in these proceedings for more details.

4.2 Conclusion

We have proposed an efficient method to maximize the likelihood of a model of noisy mixtures of stationary sources by implementing the EM algorithm on spectral domains. The procedure jointly estimates all the parameters: mixing matrix, average source spectra, noise level in each sensor. Spectral averaging offers large computational savings, especially when dealing with images.

Since the inference principle is maximum likelihood for a 'smooth' Gaussian stationary model, we expect a good statistical efficiency when the source spectra are reasonably smooth (even though we saw little performance degradation in our experiments when using a very coarse $Q = 2$ spectral partition) and when the sources actually are Gaussian. In the CMB application, some components are very close to Gaussian (the CMB itself) while others are strongly non Gaussian; it is not clear yet how to best combine non Gaussian information with spectral diversity.

As final note, we recall that, in the noise-free case, the ability to blindly separate Gaussian stationary components rests on *spectral diversity*: the spectra of any two sources should not be proportional. The noisy case is complicated by the fact that the noise parameters also have to be estimated.

Future research should cover many issues: blind identifiability in unknown noise, choice of the spectral domains, integration of non Gaussian information, integration of prior information,...

References

[1] C. Baccigalupi *et al.* Neural networks and separation of cosmic microwave background and astrophysical signals in sky maps. *MNRAS*, 318:769–780, Nov. 2000.

[2] E. Moulines *et al.*, Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. ICASSP'97*, vol. 5, pp. 3617–20, 1997.

[3] D.-T. Pham. Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion. In *Proc. EUSIPCO 2000*, pp. 621–624, 2000.

[4] H. Snoussi *et al.*, Bayesian blind component separation for cosmic microwave background observations. In *Proc. MAXENT 2001*, 2001. astro-ph/0109123.

[5] J.-F. Cardoso, Looking for components in the Universe's oldest data set. In *Proc. EUSIPCO 2002* (these proceedings).