FAST MCMC SPECTRAL MATCHING SEPARATION IN NOISY GAUSSIAN MIXTURES.
APPLICATION TO ASTROPHYSICS

Hichem Snoussi

ISTIT/M2S, University of Technology of Troyes, 12, rue Marie Curie, 10000, France
Email: name@utt.fr

ABSTRACT

The main contribution of this paper is to present an efficient Bayesian solution to the blind separation of stationary Gaussian sources in a noisy mixture. The proposed approach consists of sampling the a posteriori distribution of all the unknown parameters involved in the statistical description of the blind source separation (BSS) problem. The log-posterior criteria can be interpreted as a penalized matched filtering of covariance matrices in the Kullback-Leibler metric. The success of the blind estimation relies on the non stationarity of the second order statistics and their inter-source diversity. Drawing from the joint a posteriori distribution is implemented via Gibbs sampling based on the natural data augmentation structure of the BSS model. A key point in the fast implementation of the Gibbs sampling consists of replacing the hidden sources conditional sampling by sampling only some few matrix statistics, based on the Bartlett decomposition. The proposed solution is successfully applied to blindly separate distinct astrophysical emissions on simulations of future observations with the High Frequency Instrument of the Planck space mission, due to be launched in 2007.

1. INTRODUCTION

Since the beginning of the last decade, extensive research has been devoted to the problem of blind source separation (BSS). The attractiveness of this particular problem is essentially due to both its applicative and theoretical challenging aspects. This research has given rise to the development of many methods aiming to solve this problem. An interesting particularity of this emerging field, still open to more research, is the fact that the theoretical development evolves in pair with the real world application specificities and requirements.

In this paper, we assume that the observations are m images \((X^t)_{t=1..m}\), each image \(X^i\) is defined on a finite set of sites \(S\), corresponding to the pixels of the image: \(X^i = (x^i_r)_{r \in S}\). The observations are noisy linear mixture of \(n\) sources images \((S^i)_{i=1..n}\) defined on the same set \(S\):

\[
x^i_r = \sum_{j=1}^{n} a_{ij} s^j_r + n^i_r, \ r \in S, \ i = 1..m,
\]

where \(A = (a_{ij})\) is the unknown mixing matrix, \(N^i = (n^i_r)_{r \in S}\) is a zero-mean white Gaussian noise with variances \(\sigma^2_{n_r}\). At each site \(r \in S\), the matrix notation is:

\[
x_r = As_r + n_r.
\]

The noise and source components are assumed to be statistically independent.

The challenging aspect of the BSS problem is the absence of any exact information about the mixing matrix \(A\).

Based on i.i.d source modeling, many proposed algorithms are designed to linearly demix the observations \(x_1\), \(x_2\). The separation principle in these methods is based on the statistical independence of the reconstructed sources (Independent Component Analysis) [1]. However, ICA is designed to efficiently work in the noiseless case. In addition, with the i.i.d assumption, the separation necessarily relies on high order statistics and treating the noisy case with the maximum likelihood approach leads to complicated algorithms [2].

Discarding the i.i.d assumption, source separation can be achieved with second order statistics. For instance, second order correlation diversity in the time domain, frequency domain or time frequency domain [3–5] are successfully used to blindly separate the sources. Stationarity and non stationarity can approximately be seen as dual under Fourier transformation. For instance, based on the circular approximation, it is shown that a finite sample correlated temporal stationary signal has a Fourier transform with non stationary decorrelated samples. A maximum likelihood method was proposed in [6, 7] to separate noisy mixture of Gaussian stationary sources exploiting this temporal \(\rightarrow\) spectral duality. This method was extended to the time-frequency domain to deal with non stationary sources [8]. In this work, we propose a full Bayesian implementation exploiting this duality. The proposed algorithm provides, in the stationary regime, samples drawn from the posterior distributions of all the variables involved in the problem leading to great flexibility in the cost function choice.

2. BAYESIAN FORMULATION

Assuming the stationarity of the sources in the spatial domain, an useful approximation is to consider the covariance matrices circular. This approximation consists in circularizing the sources (see Figure 1). Then, the covariance matrices can be diagonalized in the Fourier basis with eigenvalues coinciding with the spectrum of the signal. Consequently, Fourier transforming the signals leads to a decorrelated signal with a diagonal covariance equal to the spectrum of the signal. Considering only the second order statistics, this is equivalent to a non stationary white Gaussian process.

\[
x_1 \xrightarrow{\text{1-lag correlation}} n_1 = E[x_1 x_2]
\]

Fig. 1. The circular approximation consists in circularizing the vector \(x\). For instance, the 1-lag correlation \(r_1 = E[x_1 x_2]\) is equal to \(E[x_n x_1]\).
Transforming the data by Fourier decomposition, the mixture model, at each Fourier mode \( k \) of the spectral domain, is,

\[
x(k) = A s(k) + n(k)
\]

The sources are modeled by a non-stationary white Gaussian process (real and imaginary components are independent). At each frequency \( k \), the sources \( s_k \) has a centered Gaussian distribution with diagonal covariance \( P_k = E [s_k s_k^*] \) (the diagonality reflects the sources independence),

\[
s_k \sim \mathcal{N}(0, P_k).
\]

The diagonal elements \( \sigma^2(k) = P_{ss}(k), k = 1..K \) of the matrices \( P_k \) are the power spectra of the sources. The noise is assumed to be centered white Gaussian with constant spectrum \( R_n = E [n_k n_k^*] \). We also assume that the matrix \( R_n \) is diagonal and that the diagonal elements may have different values (different noise detector levels).

Given the observed data \( X \), the \textit{a posteriori} distribution of the whole parameter \( \theta = (A, R_n, \{P_k\}_{k \in S}) \) contains all the information that we can extract from the data. According to the Bayesian rule, we have:

\[
p(\theta \mid X) \propto p(X \mid \theta) \pi(\theta) \propto \int_{S} p(X, S \mid \theta) dS \pi(\theta)
\]

where \( \pi(\theta) \) stands for the \textit{a priori} distribution of the parameters. A discussion about the selection of the prior distribution could be found in [2]. In this work, a flat (improper) prior is chosen for \( A \) and Jeffreys priors for \( R_n \) and \( P_k \). Note the natural data augmentation structure of the BSS problem which consists in integrating over the sources \( S \).

We note that the number of the power spectrum coefficients \( P_{ss} \) to be estimated is \( K \), the size of the whole sample. For this reason, we partition the spectral interval \([1..K]\) into \( L \) sub-intervals \([D_l]_{l=1}^L\) (see Figure 2). Then, we reduce the number of spectral coefficients to be estimated to \( K/L \). Consequently, the observation covariance is constant in each sub-interval: \( R_k = R_l = A P_k A^* + R_n \), \( \forall k \in D_l \). The likelihood expression can be rewritten by re-partitioning the modes, leading to the following expression of the log-posterior:

\[
\log p(\theta \mid x_{1..K}) = - \sum_{l=1}^{L} w_l D_{KL}(R_l, \tilde{R}_l) + \log \pi(\theta \mid I) \tag{1}
\]

where \( w_l = |D_l| \) is the number of modes belonging to the sub-interval \( D_l \), \( D_{KL}(\cdot, \cdot) \) is the Kullback-Leibler divergence between the structured observation covariances \( R_l \) and the empirical observation covariances \( \tilde{R}_l = \sum_{k \in D_l} x_k x_k^* \). The maximum a posteriori criteria can be considered, thus, as a penalized version of the spectral covariance matching.

### 2.1. Fast Gibbs Spectral Matching

In spite of its explicit analytic form, the posterior function (1) is difficult to sample from. However, we can make use of the hidden variable structure of the problem (the sources are the hidden variables) to implement the Gibbs sampling algorithm.

**Crude version:** The Gibbs sampler consists of sampling the parameters \( \theta \) and the sources \( S \) in a cyclic way:

1. draw the source maps \( \tilde{S}^{(k)} \sim p(S \mid X, \tilde{\theta}^{(k-1)}) \)
2. draw the parameters \( \tilde{\theta}^{(k)} \sim p(\theta \mid X, \tilde{S}^{(k)}) \)

This Bayesian sampling produces a Markov chain \( \tilde{\theta}^{(\cdot)} \), ergodic and having as stationary distribution, the marginal posterior \( p(\theta \mid X) \).

The sources are sampled according to their Gaussian posterior density. At each spectral mode \( k \) (pixel), the mean and the covariance of the source vector \( s_k \) are:

\[
\begin{align*}
\mu_k &= V_k A^* R_n^{-1} x_k \\
V_k &= (A^* R_n^{-1} A + P_k^{-1})^{-1} \tag{2}
\end{align*}
\]

**Remark 1** Note that the mean \( \mu_k \) represents the Wiener filtering of the sources. Then, we have an optimal reconstruction at each step of the algorithm.

The conditional sampling of \( \theta \) (second step) is an easy task (this represents the principal reason of introducing the hidden sources). The conditional distribution \( p(\theta \mid X, S) \) is factorized into two conditional distributions:

\[
p(\theta \mid X, S) \propto p(A, R_n | X, S) p(P | S)
\]

leading to a separate sampling of \( (A, R_n) \) and \( \{P_k\}_{k=1}^L \). Choosing the Jeffreys priors for the covariances and a flat prior for the mixing matrix, the \textit{a posteriori} distributions are:

- **Inverse Wishart** for the noise covariance and the source covariance.
- **Normal** for the mixing matrix.

The expressions of these distributions are given below:

\[
\begin{align*}
\nu_p &= 2|S| - n, \quad \Sigma_p = \frac{1}{2} |S| - n \left( R_{xx} - R_{ss} R_{ss}^{-1} R_{ss} \right)^{-1} \\
A &= \mathcal{N}(A_p; \Gamma_p), \quad A_p = R_{xx} R_{ss}^{-1}, \quad \Gamma_p = \frac{1}{|S|} R_{ss}^{-1} \otimes R_n
\end{align*}
\]

where \( Wi_{d}(R; \nu, \Sigma) \) is the Wishart distribution\(^1\) and we have defined the empirical statistics \( R_{xx} = \frac{1}{|S|} \sum_{k} x_k x_k^* \), \( R_{ss} = \frac{1}{|S|} \sum_{k} s_k s_k^* \)

\[^1\] \(Wi_{d}(R; \nu, \Sigma) \propto |R|^{(\nu-d+1)/2} \exp \left[ -\frac{1}{2} \text{Tr} \left( R \Sigma^{-1} \right) \right] \)
and \( R_{ss} = \frac{1}{|D|} \sum_{k \in D} s_k s_k^* \) (the sources \( S \) are generated in the first step of the Gibbs sampling).

Fast version: Despite the partition of the spectral domain into some few rings \( \{ D_l \} \), the sampling of the source maps at each iteration is very demanding. In fact, the mean of the \( a \ posteriori \) Gaussian \( (2) \) still depends on the mode \( k \) through the observed vector \( x_k \). In addition, the Gibbs sampling may require a great number of iterations to ensure the convergence of the Markov chain. This leads to a heavy computational algorithm. However, one can considerably reduce the computational cost without losing the algorithm performance by noting the following: In the crude implementation, the sources \( S \) are sampled but only the statistics \( R_{xs} \) and \( R_{ss} \) are used when drawing the parameters \( \theta \) (see equation \( (3) \)). The acceleration of the Gibbs sampling consists then in sampling the statistics \( R_{xs} \) and \( R_{ss} \) rather. We show in the following how drawing the statistics is easily performed.

First, define the following empirical covariances in each subdomain \( D_l \):

\[
R_{xs}^{(l)} = \frac{1}{|D_l|} \sum_{k \in D_l} x_k s_k
\]

\[
R_{ss}^{(l)} = \frac{1}{|D_l|} \sum_{k \in D_l} s_k s_k^*
\]

\[
R_{xs}^{(l)} = \frac{1}{|D_l|} \sum_{k \in D_l} x_k s_k^*
\]

As the source vector \( s_k \) in the sub-domain \( D_l \) has the following mean and covariance:

\[
\mu_k = V_l A^* R_n^{-1} x_k
\]

\[
V_l = (A^* R_n^{-1} A + P_1^{-1})^{-1}
\]

the empirical covariances \( R_{xs}^{(l)} \) and \( R_{ss}^{(l)} \) can be rewritten (after some simple algebraic manipulations) as follows:

\[
R_{xs}^{(l)} = R_{xs}^{(l)} R_n^{-1} V_l + U_{ss}^* C^*
\]

\[
R_{ss}^{(l)} = V_l A^* R_n^{-1} R_{xs}^{(l)} + C_l U_{ss} R_n^{-1} V_l + C U_{ss} C^*
\]

where \( V_l, C_l, C_l^* \) (Cholesky decomposition) and \( U_{ss} \) and \( U_{ss} \) are two random matrices, having respectively a normal and a Wishart distribution:

\[
U_{ss} \sim \mathcal{N}(0, \frac{1}{|D|} R_{xs}^{(l)} \otimes I_n)
\]

\[
U_{ss} \sim \mathcal{W}_{|D|}((2|D|)I_n)
\]

We have thus avoided the sampling of the sources and we sample rather only two random matrices \( U_{ss} \) and \( U_{ss} \) in each sub-domain \( D_l \). A key point in this implementation is the use of the Bartlett decomposition in order to sample from the Wishart distribution \([9]\). In fact, this decomposition is very effective and considerably reduces the computational cost, especially when the degree of freedom of the Wishart distribution increases. Note that this degree is proportional to the number of pixels in each sub-domain \( D_l \). The total statistics \( R_{xs} \) and \( R_{ss} \) are simply obtained by linear combination of the matrices \( R_{xs}^{(l)} \) and \( R_{ss}^{(l)} \) as follows:

\[
R_{xs} = \frac{1}{\sum_l |D_l|} \sum_{l=1}^L |D_l| R_{xs}^{(l)}
\]

\[
R_{ss} = \frac{1}{\sum_l |D_l|} \sum_{l=1}^L |D_l| R_{ss}^{(l)}
\]

3. SIMULATION RESULTS: CMB IMAGING

The proposed separating is applied to separate astrophysical emissions. The prime interest is the extraction of the Cosmic Microwave Background (CMB) component. The importance of measuring anisotropies of the (CMB) to constrain cosmological models is now well established. In the past ten years, tremendous theoretical activity demonstrated that measuring the properties of these temperature anisotropies will constrain drastically the cosmological parameters describing the matter content, the geometry, and the evolution of our Universe [10].

The simulations are based on the same conditions of \([6]\). We simulate, following the model introduced in Section 1, observations from 6 detectors \((N_d = 6)\) corresponding to the 6 HFI frequency channels. To reduce the computing time we restrict our simulations to small patches of the sky of \(300 \times 300\) square pixels of 2.5 arcmin. Working on small maps finds also a theoretical justification in the fact that the spectrum of emission of the dust (for instance) is known to vary slightly from a region of the sky to the other. The instrumental noise for each detector is considered white and Gaussian of zero mean and of rms expected for the HFI instrument channels. We also assume no correlation between the noise of the different detectors.

The simulations include 3 components \((N_c = 3)\), CMB radiation, dust emission and SZ effect in clusters of galaxies. See \([6]\) for more details about simulation models of these component maps. Spatial templates of these components are shown in figure 3. The simulated observations are obtained from the superposition of the three above spatial templates with mixing coefficients set by the electromagnetic spectra of emission and the frequency of observation. The observations used as inputs in the separation procedure (with noise added at the level expected for Planck HFI detectors) are shown in figure 4.

The proposed Fast Gibbs Spectral Matching algorithm is applied on the observed images in figure 4 to obtain the parameter Markov chains: \(\{ \hat{A}^{(h)}, \{ \hat{R}_n^{(h)} \}, \{ \hat{P}^{(h)} \}\) \(\text{after} \ h_0 \text{ iterations (warming-up), the samples} \( \{ \hat{\theta}_l^{(h_0+h)} \}\) \text{can be considered to be approximately drawn from their marginal posterior distribution} \( p(\theta_l | X) \). Then, by the ergodic theorem, one can approximate \textit{a posteriori} expectations by empirical means:

\[
\mathbb{E}[f(\theta_l) | X] \approx \frac{1}{H} \sum_{h=1}^H f(\hat{\theta}_l^{(h_0+h)})
\]

For instance, a \textit{a posteriori} expectations could be used in order to estimate the different parameters. The \textit{a posteriori} expectations are then simply approximated by the empirical means of the Markov chains. Figure 6 shows the Markov chain of the mixing coefficient \(A_{21}\) and its marginal \textit{a posteriori} distribution \( p(A_{21} | X) \). All the elements of the Markov chain are divided by the true value of the mixing coefficient in order to show the accuracy of the estimation. Note that the \textit{a posteriori} distribution is very peaky around the true value, due to the great size of the observation sample. The source components (maps) are then reconstructed with the Wiener filter based on the estimated mixing matrix, the noise levels and the source spectra. Figure 5 shows the estimated maps. Note the accuracy of the estimation for the CMB and Dust components. The estimation of the SZ component is less accurate. However, the high intensity points are recovered. A post-processing should be
performed on this component in order to recover only the high intensity points with a smoother background. Figure 7 shows the accuracy of estimation of the detector noise levels and the CMB spectrum

Fig. 3. Spatial template of the true CMB, dust and SZ components

Fig. 4. Simulated observations of the six HFI channels.

Fig. 5. Spatial template of the estimated CMB, dust and SZ components.

4. REFERENCES


