

Hidden Markov Models for Wavelet Based Blind Source Separation

Mahieddine M. Ichir and Ali Mohammad-Djafari

Abstract—In this paper we consider the problem of blind source separation (BSS) in the wavelet domain. We propose a Bayesian estimation framework for the problem where different models of the wavelet coefficients are considered: the Independent Gaussian Mixture (IGM) model, the Hidden Markov Tree (HMT) model and the Contextual Hidden Markov Field (CHMF) model. For each of the three models we give expressions of the posterior laws and propose appropriate Markov Chain Monte Carlo (MCMC) algorithms in order to perform unsupervised joint blind separation of the sources and estimation of the mixing matrix and hyper parameters of the problem.

Indeed, in order to achieve an efficient joint separation and denoising procedures in the case of high noise level in the data, a slight modification of the exposed models is presented: the Bernoulli Gaussian (BG) mixture model, which is equivalent to a hard thresholding rule in denoising problems. A number of simulations are presented in order to highlight the performances of the aforementioned approach: i) in both high and low signal to noise ratios. ii) comparing the results with respect to the choice of the wavelet basis decomposition.

I. INTRODUCTION

Blind source separation (BSS) has been an active area of research these last two decades [1], [2], [3], [4]. Many encountered problems can be reasonably viewed as being blind source separation problems [4]. One of the most developed solutions to the problem is *Independent Component Analysis* (ICA) [1], [4]. It consists mainly in finding independent components that may represent the unobserved sources. This method has assessed its performances in many applications. However, the basic ICA model does not explicitly account for any observation noise or model errors. Nevertheless, it is by far a fast method of source separation for exact instantaneous mixing and noise free models.

To account for noise or model uncertainties, higher order statistics based methods have been considered as in [5]. However, such methods do account for Gaussian noise only and suffer from outliers. Cao et al. in [6] developed a nonlinear ICA solution robust to outliers with a pre-whitening primary step that accounts for observation noise. In order to account for time structure of the source signals, extensions to basic ICA approaches have been considered: in [7], [8], [9] joint diagonalization of time delayed second order matrices have been considered whereas in [10] diagonalization of higher order statistics have been considered. This is an extension of the JADE [11] implementation of ICA, combining high order statistics to time delayed correlation matrices to be more robust to noise and to account for time coherence.

In this paper, we consider Bayesian estimation framework for the BSS problem [12], [13], [14]. Bayesian estimation is a

natural and hierarchical way of deriving posterior distributions through appropriate assignment of:

- priors (prior models for all the unknown parameters) translating any prior knowledge one may have,
- likelihood describing statistically the observational (forward) model through assumptions made on the noise or model uncertainties.

In that context, Rowe in [12] considered Gaussian priors for sources re-deriving thus the Factor Analysis (FA) solution to the problem. Snoussi et al. in [15] and Choudrey et al. in [16] considered a mixture of Gaussians prior leading to efficient joint segmentation and separation solutions for 2D BSS problems.

These methods have been considered in the direct (observations) domain in contrast to other transform based methods such as time-frequency [17] and wavelets [18], [19]. Transform domain methods rely on the fact that usually linear and invertible transforms rearrange the data, leaving them a structure simpler to model.

In this paper, we transport the BSS problem to the wavelet domain where the parsimonious property of the wavelet transform helps us to assign appropriate priors for the wavelet coefficients of the sources. Wavelet domain Bayesian blind source separation (wavelet Bayes-BSS) has already been considered in [18], [19] with generalized exponential prior models for source wavelet coefficients. These particular models present, however, some optimization difficulties.

Crouse et al. [20], in a wavelet based denoising problem, proposed to model the wavelet coefficients by a two Gaussians mixture prior model which captures efficiently the wavelet transform properties of a wide class of signals. Being a mixture of Gaussians, this prior model remains tractable (conditional linear posterior estimates) while keeping good approximation characteristics.

Based on the Gaussians mixture prior, we consider three different models for the wavelet coefficients of the unobserved sources: i) A first model assuming independence across and through the wavelet decomposition scales, the Independent Gaussians Mixture (IGM) model. ii) A second model, proposed by Crouse et al. in [20], that accounts for an inter scale correlation between the wavelet coefficients on a quad tree representation. This correlation is expressed through a first order Markov chain model, the Hidden Markov Tree (HMT) model. iii) A third prior model that we propose based on hidden Markov fields, accounts for inter and intra-scale correlations, the Contextual Hidden Markov Field (CHMF) model. It is also based on a quad tree representation. A

comparison of these three models in Bayesian BSS is also presented.

In order to be able to perform blind source separation for high noisy mixture observations, an additional constraint on the two Gaussians mixture prior distribution must be considered. In [21], [22], [23] close connections between hard/soft thresholding and wavelet based Bayesian denoising have been established in the case of generalized exponential prior distributions. Pesquet et al. in [24] established similar relations for Bernoulli-Gaussian (BG) mixture prior models. The Bernoulli-Gaussian mixture distribution is in fact no more than a limiting case of the two Gaussians mixture model presented in [20]. This will enable us to implement, with no major modification of the estimation routines developed for the two Gaussians mixture prior, an efficient joint separation denoising procedure in the case of high noise level affected observations blind source separation problems.

With the particular choice of the presented prior distributions, conditional posterior distributions of the unknown parameters (unobserved sources, mixing matrix, noise variance and hyperparameters) are explicit and particularly easy to sample. This offers the ability to implement efficient and simple Markov Chain Monte Carlo (MCMC) algorithms through Gibbs sampling for the optimization part. In that context and in order to be able to properly sample the hidden variables corresponding to the three different prior models, conditional distributions of the hidden variables have been re-derived for the wavelet tree representation: two algorithms presented in [25] for sampling 1D hidden Markov variables are extended for the 2D quad tree hidden variables of the wavelet coefficients.

This paper is organized as follows: In section II we introduce the blind source separation (BSS) problem and briefly present the main classical solutions to the problem. In section III we present the Bayesian formulation of the blind source separation (Bayes-BSS) problem and describe the prior assignment of the unknown parameters: the noise variance, the mixing matrix and the unobservable sources. In section IV-A we briefly introduce the wavelet transform used in our approach. Through a description of the main properties of the wavelet coefficients of signals (especially 2D signals) we define, in details, the different prior models we will use for the wavelet based Bayes-BSS in section IV-C, IV-D and section IV-E. The expressions of the conditional posteriors are detailed in section V for the MCMC algorithm. In section VI a simple procedure is presented in order to perform a joint source separation and denoising in the case of high noisy observations. We then conclude this work by presenting some simulation examples and comparisons in section VII and a conclusion in section VIII. Appendix A, B and C detail sampling schemes corresponding to the different prior models of section IV-C, IV-D and section IV-E.

II. CLASSICAL BSS SOLUTIONS

Blind source separation (BSS) consists of recovering unobserved sources from a set of their *linear* and *instantaneous*

mixtures, generally described by:

$$\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k) + \boldsymbol{\epsilon}(k), \quad (1)$$

where k can be a scalar index representing time, frequency, wavelength (1D cases), or a vector index representing pixels positions, time-frequency, time-scale (2D cases). In the following, we refer to k as "time" and to column vector dimension as "space". $\mathbf{x}(k)$ is the m -column vector of the observed mixtures data, $\mathbf{s}(k)$ is the n -column vector of the unobserved sources, \mathbf{A} is the $(m \times n)$ mixing matrix representing the linear and instantaneous mixing process and $\boldsymbol{\epsilon}(k)$ is the m -column vector that represents an observation noise or model error: all over this paper, it is assumed Gaussian, centered, temporarily white and spatially independent, with a covariance matrix $\mathbf{R}_{\boldsymbol{\epsilon}} = \text{diag}(\sigma_{\epsilon,1}^2, \dots, \sigma_{\epsilon,m}^2)$. The model (1) can be equivalently written in a matrix form:

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{E}, \quad (1')$$

where \mathbf{X} , \mathbf{S} and \mathbf{E} are matrices with columns respectively $\mathbf{x}(k)$, $\mathbf{s}(k)$ and $\boldsymbol{\epsilon}(k)$ for $k = 1, \dots, K$.

Classical source separation methods consider a noise free observational model of the form:

$$\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k), \quad (2)$$

and try to find, by some nonlinear optimization criteria, a *separating* matrix \mathbf{B} (generally an estimation of the inverse of \mathbf{A} up to a permutation \mathbf{P} and a scale indeterminacy \mathbf{D} : $\mathbf{B} = \mathbf{P}\mathbf{D}\mathbf{A}^{-1}$). The sources are then estimated by:

$$\mathbf{y}(k) = \mathbf{B}\mathbf{x}(k). \quad (3)$$

A. Principal Component Analysis (PCA)

If we consider second order stationary sources $\mathbf{s}(k) \sim \mathcal{N}(0, \mathbb{I}_n)$, $\forall k$, the distribution of the observations $\mathbf{x}(k)$ according to the mixing model (2) is $\mathcal{N}(0, \boldsymbol{\Sigma}_x = \mathbf{A}\mathbf{A}')$ and the distribution of $\mathbf{y}(k)$ is $\mathcal{N}(0, \mathbf{B}\boldsymbol{\Sigma}_x\mathbf{B}')$. Since $\mathbf{y}(k) = \mathbf{P}\mathbf{D}\mathbf{s}(k)$ then $\mathbf{B}\boldsymbol{\Sigma}_x\mathbf{B}' = \mathbb{I}_n$ and a possible solution is:

$$\mathbf{B} = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}^\dagger, \quad (4)$$

where $(\mathbf{U}, \boldsymbol{\Lambda})$ are obtained by singular value decomposition (SVD) of $\boldsymbol{\Sigma}_x$. The PCA algorithm then starts by estimating $\boldsymbol{\Sigma}_x$ from the observed data and then computing \mathbf{B} using the SVD. The principal component are then obtained by (3).

B. Independent Component Analysis (ICA)

ICA can be defined as the process of decomposing the observations into mutually independent components. A fundamental measure of independence (to be minimized with respect to \mathbf{B}) is the mutual information given by:

$$\begin{aligned} I(\mathbf{y}) &= \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{\prod_i p(y_i)} d\mathbf{y} \\ &= -H(\mathbf{y}) + \sum_i H(y_i), \end{aligned} \quad (5)$$

where $H(\cdot)$ is the differential entropy. Mutual information can be equivalently written as:

$$I(\mathbf{y}) = J(\mathbf{y}) - \sum_i J(y_i) + \frac{1}{2} \log \frac{\prod_i \boldsymbol{\Sigma}_y(i, i)}{|\boldsymbol{\Sigma}_y|}, \quad (6)$$

where Σ_y is the covariance matrix of \mathbf{y} and $J(\cdot)$ is the *negentropy* which measures the distance of a distribution to the Gaussian one. ICA based methods consist, generally, on approximations of $I(\mathbf{y})$ (or equivalently $J(y_i)$) by high order cumulants [1] or nonlinear functions [26].

C. Maximum likelihood source separation

The maximum likelihood solution to BSS begins by writing the probability distribution of the observations. The log likelihood is given by:

$$\mathcal{L}(\mathbf{B}) = K \ln |\mathbf{B}| + \sum_i \sum_k \ln p(y_i). \quad (7)$$

Asymptotically, equation (7) reduces to:

$$\begin{aligned} \lim_{K \rightarrow \infty} \frac{1}{K} \mathcal{L}(\mathbf{B}) &= -\ln |\mathbf{B}| + \sum_i^n \mathbf{E}[\ln p(y_i)] \\ &= -I(\mathbf{y}) - H(\mathbf{x}). \end{aligned} \quad (8)$$

Given that $H(\mathbf{x})$ is constant, maximizing the likelihood is equivalent to minimizing the mutual information given by equation (5).

D. Time structure ICA

The ICA approach to the BSS problem has been further extended in order to account for time evolution of the original sources: in [17], the source separation problem has been considered in the time-frequency domain (with the short time Fourier transform) in order to account for time non-stationarity. In [8], [7] and [9], joint diagonalization of time delayed second order matrices have been considered in order to find the separating (orthogonal) matrix. The algorithms developed in [7] and [9] where respectively named SOBI (Second Order Blind Identification) and TDSEP (Temporal Decorrelation Source SEPARation). As a further extension of the JADE algorithm [11] (for Joint Approximate Diagonalization of Eigen matrices based on high order statistics: cumulants), Müller in [10] considered a combination of the JADE for high order statistics and the TDSEP algorithm for second order correlations to develop the JADE_{TD} algorithm: an efficient algorithm for BSS accounting for noise in the observations.

III. BAYESIAN BLIND SOURCE SEPARATION (BAYES-BSS)

In a Bayesian estimation framework, we begin by writing the posterior distribution of all the unknown parameters corresponding to the BSS problem of equation (1):

$$\begin{aligned} p(\mathbf{S}, \mathbf{A}, \mathbf{R}_\epsilon, \boldsymbol{\theta} | \mathbf{X}) \\ \propto p(\mathbf{X} | \mathbf{S}, \mathbf{A}, \mathbf{R}_\epsilon) \pi(\mathbf{S}, \mathbf{A}, \mathbf{R}_\epsilon | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}), \end{aligned} \quad (9)$$

where $p(\mathbf{X} | \mathbf{S}, \mathbf{A}, \mathbf{R}_\epsilon)$ is the likelihood function which, under an independent and identically distributed (i.i.d.) Gaussian noise, is given by:

$$p(\mathbf{X} | \mathbf{S}, \mathbf{A}, \mathbf{R}_\epsilon) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}(k) | \mathbf{A} \mathbf{s}(k), \mathbf{R}_\epsilon). \quad (10)$$

$\pi(\mathbf{S}, \mathbf{A}, \mathbf{R}_\epsilon | \boldsymbol{\theta})$ is the prior distribution and $\boldsymbol{\theta}$ represents the parameters needed to properly define the priors, commonly called *hyperparameters*.

One of the most important steps in Bayesian estimation consists of an appropriate assignment of this prior distribution. We will first assume, that the parameters of interest are *a priori independent*:

$$\pi(\mathbf{S}, \mathbf{A}, \mathbf{R}_\epsilon | \boldsymbol{\theta}) = \pi(\mathbf{S} | \boldsymbol{\theta}_s) \pi(\mathbf{A} | \boldsymbol{\theta}_A) \pi(\mathbf{R}_\epsilon | \boldsymbol{\theta}_\epsilon). \quad (11)$$

On the set of hyperparameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_s, \boldsymbol{\theta}_A, \boldsymbol{\theta}_\epsilon]$, only $\boldsymbol{\theta}_s$ will be inferred, the other hyperparameters set $\{\boldsymbol{\theta}_A, \boldsymbol{\theta}_\epsilon\}$ will be fixed once for all, reducing the number of unknown variables.

A. Noise variance prior distribution $\pi(\mathbf{R}_\epsilon | \boldsymbol{\theta}_\epsilon)$

A conjugate *Inverse Gamma* prior distribution is chosen for scale parameters [27]. The Inverse Gamma pdf is given by:

$$\mathcal{G}^{-1}(x) \propto \frac{1}{x^{\nu+1}} \exp\left(-\frac{1}{\theta x}\right) \mathbb{I}_{[0, +\infty[}, \quad (12)$$

having as its limiting distribution ($\nu = 0, \theta \rightarrow \infty$) the *non informative Jeffrey's prior*: $\pi(x) \propto 1/x$.

B. Mixing matrix prior distribution $\pi(\mathbf{A} | \boldsymbol{\theta}_A)$

The prior distribution of the mixing matrix can be described by the physical system inherent in the mixing process (translating positivity, discrete state information ...). In this paper, the elements of the mixing matrix are considered a priori Gaussian and independent:

$$\pi(a_{i,j}) = \mathcal{N}(\mu_{i,j}^a, \sigma_a^2), \quad (13)$$

where $(i, j) \in \{1, \dots, m\} \otimes \{1, \dots, n\}$.

C. Sources modeling and $\pi(\mathbf{S} | \boldsymbol{\theta}_s)$

Sources prior distribution is clearly an important step in a Bayesian solution to the BSS problem. Different models can be considered:

1. The simplest ones are the temporal i.i.d. models of the form:

$$\pi(s_i(1), \dots, s_i(K) | \boldsymbol{\theta}_s) = \prod_k \pi(s_i(k) | \boldsymbol{\theta}_s), \quad (14)$$

with $\pi(s_i(k))$ either Gaussian (linear models as for the PCA solution), or non Gaussian models, for instance the generalized p-Gaussian distributions given by:

$$\pi(s_i(k)) \propto \exp(-\gamma |s_i(k)|^p), \quad 0 < p < 2 \quad (15)$$

already considered in some wavelet based BSS as in [18], [28].

Mixture of L distributions of the form:

$$\pi(s_i(k) | p_{1,\dots,L}^i, \theta_{1,\dots,L}^i) = \sum_{l=1}^L p_l^i f_l(s_i(k) | \theta_l^i), \quad (16)$$

with $\sum_l p_l^i = 1$, has also been considered as in [14], [29], [30], [15] with variational approximations in [14], [29], [30] leading to efficient Bayes-BSS algorithms. Interpreting the

weights p_l^i as probabilities of *hidden* states z_k^i associated to the samples $s_k(i)$, we can rewrite equation (16) as:

$$\begin{aligned} \pi(s_i(k)|z_k^i, \theta_{1,\dots,L}^i) \\ = \sum_{l=1}^L p(z_k^i = l) f_l(s_i(k)|\theta_l^i) \end{aligned} \quad (17)$$

2. Time non stationarity can also be considered to enhance the i.i.d models such as second order models of the form $\mathcal{N}(\mu_k^i, \sigma_k^i)$ where the time variation of (μ_k^i, σ_k^i) has to be described [31], [32], [33].

Another interesting non stationary models widely considered in statistical signal processing are the hidden Markov models (HMM) given by:

$$\begin{aligned} \pi(s_i(k)|z_{k' \in \nu(k)}^i, \theta_{1,\dots,L}^i) = \\ \sum_{l=1}^L p(z_k^i = l | z_{k' \in \nu(k)}^i) f_l(s_i(k)|z_k^i = l, \theta_l^i), \end{aligned} \quad (18)$$

where z_k^i 's are discrete random variables taking values in the set $\in \{1, \dots, L\}$. The Markovian property is, in a general manner, expressed by $p(z_k^i | z_{k' \in \nu(k)}^i)$, where $\nu(k)$ is a neighborhood system of k . In [15], [34], [16], the f_l^i 's were taken to be Gaussian densities for a 2D Bayes-BSS problem. Indeed, such prior models enabled to have some joint segmentation and separation of the images, where the 2D gray scale images are naturally segmented into L statistically distinct regions via the *label* (hidden) variables z_k^i 's.

IV. WAVELET DOMAIN STATISTICAL SIGNAL PROCESSING

Mixture densities of equation (17) have been extensively used in statistical signal processing. However, in the direct domain, the number of density functions (L) in the prior Gaussians mixture model of equation (17) should be chosen sufficiently large in order to achieve good approximation properties. Another approach is to consider linear transforms having some particular properties that rearrange the data leaving them a structure simpler to model.

The wavelet transform is a sparse representation that transforms the data rendering them interestingly simple to model: It results, for a wide class of signals, into a *large* number of *small* coefficients and a *small* number of *large* coefficients. This has been our main motivation to adopt simple models for the wavelet coefficients in a Bayesian source separation problem: a two Gaussians mixture distribution ($L = 2$) with a hidden Markov modeling. In the following, we briefly detail the wavelet transform used and then we point out to the main properties of the wavelet coefficients that justify the choice of simple to more complex prior models.

A. The wavelet transform

Wavelets has emerged as an interesting tool for signal processing in the last two decades¹. The wavelet transform belongs to the large family of time-frequency analysis. Its

particularity is that it analyses signals over variable shape Heisenberg boxes on the time-frequency plane [36]. It is given by:

$$W_f(s, u) = \langle f(t), \psi_{s,u}(t) \rangle = \int f(t) \psi_{s,u}(t) dt, \quad (19)$$

where $\psi_{s,u}(t) = \frac{1}{\sqrt{s}} \psi(\frac{t-u}{s})$, $\|\psi(t)\| = 1$ and $\int t^p \psi(t) dt = 0$ for $p = 0, \dots, P$; P being the number of vanishing moments of $\psi(t)$. We will not go deeper in the wavelet theory, the reader could refer to [37], [36] for a detailed literature on the subject. For a discrete time signal $f[k]$, $k = 1, \dots, K$, the fast wavelet transform (FWT) [38] algorithm is given by:

$$\begin{aligned} a_j[k] &= (a_{j-1} \otimes \bar{h})[2k-1], \\ d_j[k] &= (a_{j-1} \otimes \bar{g})[2k-1] \end{aligned} \quad (20)$$

with $\bar{h}[k] = h[-k]$, $\bar{g}[k] = g[-k]$, $k = 1, \dots, 2^{-j}K$ and $j = 1, \dots, J \leq \log_2(K)$. (h, g) is a pair of quadratic mirror filters, (a_j, d_j) are respectively the scale and detail coefficients of $f[k] = a_0[k]$. For 2D images, the FWT algorithm is derived from 2D separable wavelet functions along the lines and columns of the image [38].

The FWT used in this work, is an orthonormal (dyadic) multi-resolution analysis that conserves an unchanged global number of samples. The FWT is not shift invariant, but this particular property is not important for the BSS problem we are dealing with.

The wavelet transform presents several properties of great importance in signal processing. The first one we will be interested with is:

P1. *Linearity and Inversibility*: the wavelet transform is a linear and invertible transform, more over, it is orthonormal.

This property is essential, since it implies that linear problems described by $x_{1,\dots,K} = \mathcal{K}.s_{1,\dots,K} + \epsilon_{1,\dots,K}$, where \mathcal{K} is a linear operator (convolution or point wise operator), can be equivalently described in the transform domain as $w_x^\lambda = \mathcal{W}^\dagger \mathcal{K} \mathcal{W} w_s^\lambda + w_\epsilon^\lambda = \mathcal{K}'. w_s^\lambda + w_\epsilon^\lambda$, where \mathcal{W} is the wavelet transform operator (\mathcal{W}^\dagger is the adjoint), w_x^λ is the $\lambda^{\text{th}} = (j, k_j)$ wavelet coefficient of $x_{1,\dots,K}$. The instantaneous and linear BSS problem of equation (1) can thus be equivalently described in the wavelet domain by:

$$w_x^\lambda = \mathbf{A} w_s^\lambda + w_\epsilon^\lambda, \quad \lambda = (j, k_j) \quad (21)$$

for $j = 1, \dots, J$ and $k_j = 1, \dots, 2^{-j}K$, or in a matrix form $\mathbf{W}_x = \mathbf{A} \mathbf{W}_s + \mathbf{W}_\epsilon$. In this last equation, the index λ means that the BSS of equation (1) is rewritten in each wavelet sub-band. Rewriting now the joint posterior (9) as given by equation (22), where we will infer, from now on, on the wavelet coefficients of the unobservable sources \mathbf{W}_s rather than on the sources \mathbf{S} themselves. The prior distributions of the noise variance and the mixing matrix have been already defined in sections III-A and III-B.

B. Wavelet domain prior models and $\pi(\mathbf{W}_s | \theta_s)$

A second property of the wavelet transform that is the basis of the wavelet based compression algorithms (JPEG2000 [39]) and signal approximation [40] of piecewise regular signals² is:

¹the first use of *wavelets* dates back to Haar [35]

²such signals are said to belong to the Besov space.

$$p(\mathbf{W}_s, \mathbf{A}, \mathbf{R}_\epsilon, \boldsymbol{\theta} | \mathbf{W}_x) \propto p(\mathbf{W}_x | \mathbf{A}, \mathbf{W}_s, \mathbf{R}_\epsilon) \pi(\mathbf{W}_s | \boldsymbol{\theta}_s) \pi(\mathbf{A} | \boldsymbol{\theta}_A) \pi(\mathbf{R}_\epsilon | \boldsymbol{\theta}_\epsilon) \pi(\boldsymbol{\theta}_s), \quad (22)$$

P2. *Locality and Compression*: a wavelet atom is localized both in time and frequency, the wavelet coefficients contain local information of the signals: the wavelet transform of a wide class of signals is parsimonious.

In other words, it states that many signals can be well approximated by a small number of their wavelet coefficients. This parsimonious property can be statistically modeled by *centered, peaky* and *heavy tailed* distributions.

One such possible prior model is the generalized exponential family (gpG) [38], [22] given by equation (15). This particular model allowed to establish close connections between wavelet hard/soft thresholding [41] and Bayesian estimation [21], [22], [23]. However the gpG prior model results in non linear optimization problems which are not trivial.

A second possible prior model proposed by Crouse et al. in [20] is a two Gaussians mixture prior distribution:

$$p(w_s^\lambda | p_L, \tau_L, \tau_H) = p_L \mathcal{N}(w_s^\lambda | 0, \tau_L) + (1 - p_L) \mathcal{N}(w_s^\lambda | 0, \tau_H), \quad (23)$$

with $\tau_L \ll \tau_H$, where $p_L = \text{Prob}\{\text{wavelet coefficient} \in \text{Low energy state}\}$, and $p_H = 1 - p_L = \text{Prob}\{\text{wavelet coefficient} \in \text{High energy state}\}$. Note that the set $\{\tau_L, \tau_H, p_L\}$ is proper to each sub-band.

The two Gaussians mixture model presents the advantages of modeling efficiently the wavelet coefficients with only three parameters (p_L, τ_L, τ_H) as compared to the direct space Gaussians mixture model parametrized by $(3L - 1)$ parameters ($p_l's, \mu_l's, \tau_l's$).

C. Independent Gaussians mixture model (IGM)

Based on the locality property (P2), the wavelet coefficients are often considered independent within and across scales, leading to simple but fairly efficient algorithms:

$$p(z_\lambda) = \prod_j \prod_{k_j} p(z_\lambda = q), \quad \lambda = (j, k_j), \quad (24a)$$

$$\begin{aligned} p(w_s^\lambda | z_\lambda) &= \prod_j \prod_{k_j} p(w_s^\lambda | z_\lambda = q) \\ &= \prod_\lambda \mathcal{N}(w_s^\lambda | 0, \tau_q) \end{aligned} \quad (24b)$$

with $q \in \{L, H\}$ and $\Lambda = \cup_{j=1}^J \{j; k_j = 1, \dots, 2^{-j}K\}$. In Bayesian estimation, we need mainly to write the posterior distribution $p(z_\lambda = q | w_s^\lambda, p_q)$, where $p_q = p(z_\lambda = q)$. Appendix A describes a Gibbs based MCMC algorithm for sampling the hidden variables z_λ for the IGM model and the associated parameter p_q .

Notations

In order to go further in the description of the two following prior models, some notations have to be fixed in conjunction with Fig. 1:

- 1) w_θ^λ denotes the k_j^{th} wavelet coefficient of $\theta_{k=1, \dots, K}$ at resolution j , with $\lambda = (j, k_j)$ where $j = 1, \dots, J$ and $k_j = 1, \dots, 2^{-j}K$ represents a given node on the graph. z_λ denotes a *binary* random variable associated to w_θ^λ .
- 2) $T_{(J, k_J)}$ denotes the likelihood wavelet tree from the root node (J, k_J) to the leaf nodes $(1, k_1's)$. T_λ denotes the likelihood wavelet subtree from the node $\lambda = (j, k_j)$ down to its leaves. $T_{\setminus \lambda}$ denotes the likelihood wavelet subtree from root nodes $\lambda = (J, K_J)$ down to $\lambda = (j + 1, k_{j+1})$.
- 3) \mathcal{C}_λ denotes the set of the direct descendant nodes (children) of node λ , \mathcal{P}_λ its direct ascendant (parent) node and ν_λ the set of its neighboring nodes (in this work we consider only a first order neighboring system).

D. Hidden Markov tree model (HMT)

The main limitations of the IGM model, presented earlier, is that it lacks local correlations. Crouse et al. in [20] proposed a novel model in order to account for inter scale correlations based on an additional property of the wavelet coefficients:

P3. *Persistence*: the wavelet coefficients propagates across scales.

A homogeneous Markov chain model is then defined to statistically describe this particular property:

$$\begin{aligned} p(z_\lambda = q) &= \sum_{q'} p(z_\lambda = q | z_{\mathcal{P}_\lambda} = q') p(z_{\mathcal{P}_\lambda} = q'), \\ &= \sum_{q'} \pi_{q'q} p(z_{\mathcal{P}_\lambda} = q') \end{aligned} \quad (25)$$

with $\{q, q'\} \in \{L, H\}$. $\pi_{q'q}$ is the transition probability from the parent node \mathcal{P}_λ to the node λ . The likelihood function of the sources wavelet coefficients is similarly given by equation (24b). Appendix B.1 and B.2 detail two possible MCMC algorithms for sampling the HMT variables:

- *global updating* (Appendix B.1): describes an iterative method based on the forward backward formula [42] designed to determine posterior marginals of the hidden variables for 1D Markov chains. The forward and backward equations are rewritten for the quad tree model of Fig. 1 and sampling distributions are detailed.
- *local updating* (Appendix B.2): describes a Gibbs type sampling algorithm based on a procedure proposed in [25] to sample 1D Markov chains hidden variables from posterior conditionals. These equations have been also rewritten to match the quad tree representation of Fig. 1.

E. Contextual hidden Markov field model (CHMF)

A fourth property of the wavelet coefficients in conjunction with the aforementioned properties, allow us to propose a contextual hidden Markov field (CHMF) model for the wavelet coefficients in order to jointly account for inter and intra scale correlations.

P4. *Clustering*: the wavelet coefficients are locally correlated: if a wavelet coefficient is large/small, then its neighboring coefficients are likely to be large/small.

Intra and inter scale correlation can be statistically described by:

$$\begin{aligned} \pi(z_\lambda = q | z_{\nu_\lambda}, z_{\mathcal{C}_\lambda}) \\ \propto \exp\left(\beta_1 \sum_{r \in \nu_\lambda} \delta_{(z_r=q)} + \beta_2 \sum_{r \in \mathcal{C}_\lambda} \delta_{(z_r=q)}\right), \end{aligned} \quad (26)$$

where β_1, β_2 are some predefined constants. The likelihood function is similarly given by equation (24b). This model, in contrast to the HMT model of section IV-D, links a given node $\lambda = (j, k_j)$ to its direct descendants (children) $\mathcal{C}_\lambda = \bigcup_{k_{j-1} \in \mathcal{C}_\lambda} \{(j-1, k_{j-1})\}$, allowing efficient estimates of the model parameters at the finest resolution with $K_1 = K/2$ data samples. The CHMF model defined on discrete variables having only two states (binary random variables) is commonly known as the Ising model. Appendix C details the Gibbs based MCMC algorithm for sampling the CHMF hidden variables.

V. WAVELET BASED BAYES-BSS AND MCMC SAMPLING

Now that we have properly defined appropriate prior models (prior distributions) for the wavelet coefficients, the posterior distribution of equation (22) is rewritten as in equation (27), where the introduced (hidden) variable \mathbf{Z} and its corresponding prior $\pi(\mathbf{Z}|\theta_s)$ have been defined in sections IV-C, IV-D and IV-E and the prior distribution $\pi(\mathbf{W}_s|\mathbf{Z}, \theta_s)$ is given by equation (24b). Markov Chain Monte Carlo methods will allow us to generate samples from the posterior (27) and then estimate the posterior mean by its corresponding empirical mean. A Gibbs sampler, as described below:

At iteration t :

$$\begin{cases} \mathbf{Z}^t & \sim p(\mathbf{Z}|\mathbf{W}_x, \mathbf{A}^{t-1}, \mathbf{R}_\epsilon^{t-1}, \theta_s^{t-1}), \\ \mathbf{W}_s^t & \sim p(\mathbf{W}_s|\mathbf{W}_x, \mathbf{Z}^t, \mathbf{A}^{t-1}, \mathbf{R}_\epsilon^{t-1}, \theta_s^{t-1}), \\ \mathbf{A}^t & \sim p(\mathbf{A}|\mathbf{W}_x, \mathbf{W}_s^t, \mathbf{R}_\epsilon^{t-1}, \theta_A), \\ \mathbf{R}_\epsilon^t & \sim p(\mathbf{R}_\epsilon|\mathbf{W}_x, \mathbf{W}_s^t, \mathbf{A}^t, \theta_\epsilon), \\ \theta_s^t & \sim p(\theta_s|\mathbf{W}_s^t, \mathbf{Z}^t) \end{cases} \quad (A1)$$

allows to generate samples from the respective conditional distributions. The expressions of these conditionals are detailed in the following.

A. Conditional distribution of the label variables \mathbf{Z}

The conditional distribution of \mathbf{Z} is given by:

$$\begin{aligned} p(z_\lambda = q | \mathbf{w}_x^\lambda, \theta) \\ \propto \mathcal{N}(\mathbf{w}_x^\lambda | 0, \mathbf{R}_{x|z}) \pi(z_\lambda = q | \theta_s), \end{aligned} \quad (28)$$

where $\lambda = (j, k_j)$. The vector $\mathbf{z}_\lambda = [z_\lambda^{(1)}, \dots, z_\lambda^{(n)}]$ denotes the vector of the label variables at node λ of each source (number of sources is n). The vector \mathbf{q} denotes all the possible states (2^n states) that \mathbf{z}_λ can take and $\pi(z_\lambda = \mathbf{q} | \theta_s) = \prod_{i=1}^n \pi(z_\lambda^{(i)} = q^{(i)} | \theta_s)$. The expression of $\pi(z_\lambda^{(i)} = q^{(i)} | \theta_s)$ is proper to each model. Detailed expressions of the sampling distributions are given in the Appendix. The matrix $\mathbf{R}_{x|z} = \mathbf{A}\mathbf{R}_z\mathbf{A}' + \mathbf{R}_\epsilon$ where $\mathbf{R}_z = \text{diag}(\tau_q^{(1)}, \dots, \tau_q^{(n)})$, $\tau_q^{(i)}$ being

to appear on the IEEE Transactions on Image Processing

the q^{th} Gaussian variance of wavelet coefficients of source i at resolution j .

B. Conditional distribution of the sources wavelet coefficients \mathbf{W}_s

The conditional distribution of the wavelet coefficients of the sources is given by:

$$\begin{aligned} p(\mathbf{w}_s^\lambda | \mathbf{w}_x^\lambda, \mathbf{A}, \mathbf{z}_\lambda, \theta_s) \\ \propto \mathcal{N}(\mathbf{w}_x^\lambda | \mathbf{A}\mathbf{w}_s^\lambda, \mathbf{R}_\epsilon) \mathcal{N}(\mathbf{w}_s^\lambda | 0, \mathbf{R}_z) \\ \propto \mathcal{N}(\mathbf{w}_s^\lambda | \boldsymbol{\mu}_{s|z}, \mathbf{R}_{s|z}), \end{aligned} \quad (29)$$

where $\mathbf{R}_{s|z} = (\mathbf{A}'\mathbf{R}_\epsilon^{-1}\mathbf{A} + \mathbf{R}_z^{-1})^{-1}$ and $\boldsymbol{\mu}_{s|z} = \mathbf{R}_{s|z}\mathbf{A}'\mathbf{R}_\epsilon^{-1}\mathbf{w}_x^\lambda$.

C. Conditional distribution of the mixing matrix \mathbf{A}

The conditional distribution of \mathbf{A} is given by:

$$\begin{aligned} p(\text{Vect}(\mathbf{A}) | \mathbf{W}_x, \mathbf{W}_s, \theta_A) \\ \propto p(\mathbf{W}_x | \mathbf{A}, \mathbf{W}_s, \mathbf{R}_\epsilon) \pi(\text{Vect}(\mathbf{A}) | \theta_A), \\ \propto \mathcal{N}(\boldsymbol{\mu}_{A|S}, \mathbf{R}_{A|S}), \end{aligned} \quad (30)$$

where $\mathbf{R}_{A|S} = (\mathbf{R}_\epsilon^{-1} \otimes \mathbf{C}_{ss} + \mathbb{I}_m \otimes \mathbf{R}_A^{-1})^{-1}$, $\boldsymbol{\mu}_{A|S} = \mathbf{R}_{A|S}((\mathbf{R}_\epsilon^{-1} \otimes \mathbb{I}_n) \text{Vect}[(\mathbf{C}_{xs}) + \boldsymbol{\mu}_A])$, $\mathbf{C}_{ss} = \mathbf{W}_s\mathbf{W}_s^\dagger$ and $\mathbf{C}_{xs} = \mathbf{W}_x\mathbf{W}_s^\dagger$, and $\text{Vect}[\cdot]$ is a row-wise column vectorization defined as:

$$\text{Vect} \left(\begin{bmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \dots & a_{m,n} \end{bmatrix} \right) = [a_{1,1}, a_{1,2}, \dots, a_{m,n}]'$$

D. Conditional distribution of the scale parameters $(\mathbf{R}_\epsilon, \tau_q^{(i)})$

The scale parameters represent actually the noise variances $\mathbf{R}_\epsilon = \text{diag}(\sigma_{\epsilon,1}^2, \dots, \sigma_{\epsilon,m}^2)$ and the sources wavelet coefficients variances $\{\tau_q^{(1)}, \dots, \tau_q^{(n)}\}$, with $q \in \{L, H\}$. Their respective conditional distributions are given by equation (31) and (32).

VI. BERNOULLI GAUSSIAN MODEL: JOINT SEPARATION AND DENOISING

In order to achieve efficient separation in the case of high noisy observations, additional prior information has to be considered: *A wide class of signals can be well approximated by only a few number of their wavelet coefficients*. Connections between hard/soft thresholding and Bayesian estimation have been established in [21], [24], [22] for generalized p-Gaussian (gpG) priors and for the Bernoulli-Gaussian (BG) priors in [23]. The BG is in fact a limiting case of the two Gaussians mixture:

$$\begin{aligned} \pi_{BG}(w_s) &= \lim_{\tau_L \rightarrow 0} [p_L \mathcal{N}(w_s | 0, \tau_L) + p_H \mathcal{N}(w_s | 0, \tau_H)] \\ &= p_L \delta(w_s) + p_H \mathcal{N}(w_s | 0, \tau_H), \end{aligned} \quad (33)$$

with $\tau_H \gg 0$, $\delta(\cdot)$ is the usual delta function. The MAP (Maximum A Posteriori) estimator in the case of a two Gaussians mixture prior model for a simple denoising problem: $w_x^\lambda = w_s^\lambda + w_\epsilon^\lambda$, is given by:

$$\hat{w}_{s|z}^\lambda = \begin{cases} \hat{\sigma}_L^2 / \sigma_\epsilon^2 w_x^\lambda, & \text{for } z_\lambda = L, \\ \hat{\sigma}_H^2 / \sigma_\epsilon^2 w_x^\lambda, & \text{for } z_\lambda = H \end{cases} \quad (34)$$

$$p(\mathbf{W}_s, \mathbf{A}, \mathbf{R}_\epsilon, \boldsymbol{\theta}, \mathbf{Z} | \mathbf{W}_x) \propto p(\mathbf{W}_x | \mathbf{A}, \mathbf{W}_s, \mathbf{R}_\epsilon) \pi(\mathbf{W}_s | \mathbf{Z}, \boldsymbol{\theta}_s) \pi(\mathbf{Z} | \boldsymbol{\theta}_s) \pi(\mathbf{A} | \boldsymbol{\theta}_A) \pi(\mathbf{R}_\epsilon | \boldsymbol{\theta}_\epsilon) \pi(\boldsymbol{\theta}_s), \quad (27)$$

$$\begin{aligned} p(\sigma_{\epsilon,i}^2 | \mathbf{W}_{x_i}, [\mathbf{A}\mathbf{W}_s]_i) &\propto p(\mathbf{W}_{x_i} | [\mathbf{A}\mathbf{W}_s]_i, \sigma_{\epsilon,i}^2) \pi(\sigma_{\epsilon,i}^2) \propto \left(\prod_{\lambda} \mathcal{N}(w_{x_i}^\lambda | [\mathbf{A}\mathbf{w}_s^\lambda]_i, \sigma_{\epsilon,i}^2) \right) \mathcal{G}^{-1}(\sigma_{\epsilon,i}^2 | \nu_0, \theta_0), \\ &\propto \mathcal{G}^{-1}(\nu_\epsilon, \theta_\epsilon), \quad i = 1, \dots, m \end{aligned} \quad (31)$$

where $\nu_\epsilon = K/2 + \nu_0$, and

$$\theta_\epsilon^{-1} = \frac{1}{2} \sum_{\lambda} (w_{x_i}^\lambda - [\mathbf{A}\mathbf{w}_s^\lambda]_i)^2 + \theta_0^{-1}.$$

$$\begin{aligned} p(\tau_q^{(i)} | \mathbf{W}_{s_i}, z_\lambda^{(i)} = q) &\propto p(\mathbf{W}_{s_i} | z_\lambda^{(i)} = q, \tau_q^{(i)}) \pi(\tau_q^{(i)}) \propto \left(\prod_{\lambda | z_\lambda^{(i)} = q} \mathcal{N}(w_{s_i}^\lambda | 0, \tau_q^{(i)}) \right) \mathcal{G}^{-1}(\tau_q^{(i)} | \nu_0, \theta_0), \\ &\propto \mathcal{G}^{-1}(\tau_q^{(i)} | \nu_s, \theta_s), \quad i = 1, \dots, n \end{aligned} \quad (32)$$

where $\nu_s = K_j/2 + \nu_0$, and

$$\theta_s^{-1} = \frac{1}{2} \sum_{\lambda | z_\lambda^{(i)} = q} [w_{s_i}^\lambda]^2 + \theta_0^{-1}.$$

with $1/\hat{\sigma}_{L|H}^2 = 1/\tau_{L|H} + 1/\sigma_\epsilon^2$. It defines, in a general manner, a nonlinear function of the data. For the BG prior model, the MAP estimator (34) rewrites:

$$\hat{w}_{s|z}^\lambda = \begin{cases} 0, & \text{for } z_\lambda = L, \\ \hat{\sigma}_H^2 / \sigma_\epsilon^2 w_x^\lambda, & \text{for } z_\lambda = H \end{cases} \quad (35)$$

defining clearly a hard thresholding rule. Therefore, in the presence of highly noisy observations, a BG prior model is adopted in order to perform joint source separation and denoising: the hidden variables z_λ are a posteriori sampled from their posterior probabilities, however only the high energy wavelet coefficients of the unknown sources (corresponding to $z_\lambda = H$) are sampled from their conditional posteriors while the low energy coefficients (corresponding to $z_\lambda = L$) are set to zero. The presented models (IGM, HMT and CHMF) are equivalently described in that case and no modifications are needed for posterior sampling procedures of the high energy coefficients.

VII. SIMULATION EXAMPLES

In order to highlight the performances of the proposed approach, simulation examples have been performed with data sets having different statistical characteristics. The data are sets of 256×256 gray scale pixel images.

For each example, the observations are decomposed on the ‘‘Symmlet’’ wavelets with 6 vanishing moments (these wavelets are highly symmetrical). The (256×256) observed images are decomposed up to the 3rd scale (resulting, at the coarsest resolution, to (8×8) scale pixel images = 64 data samples which seems to be a good compromise for estimation purposes at lower resolutions). For the simulation results, the Markov Chain Monte Carlo runs are given a sufficient running time in order to reach convergence (convergence in law, i.e.

the samples are, at convergence, generated from the stationary posterior distribution of equation (27)).

As an indication of performance, we give the performance index (PI) defined in [43] and the correlation coefficient matrix defined by:

$$\rho_{(x,y)}[i, j] = \frac{\langle \mathbf{x}[i], \mathbf{y}[j] \rangle}{\sqrt{\|\mathbf{x}[i]\|_2 \cdot \|\mathbf{y}[j]\|_2}} \quad (36)$$

where $\langle \cdot, \cdot \rangle$ is the usual scalar product and $\|\cdot\|_2$ is the L_2 norm. The diagonal elements of this matrix measures the correlation coefficient between the estimates to the original sources (ideally equal to unity), while the off-diagonal elements measures the inter-correlation coefficients (not necessarily equal to zero). Table I summarizes the values of this matrix for the three presented examples of Fig. 2, Fig. 5 and Fig. 7. The obtained results are compared to a time structure ICA based algorithm: the TDSEP algorithm presented in [9] and available at ‘‘<http://wwwold.first.fhg.de/~ziehe/download.html>’’.

In a first example (Fig. 2-a), two synthetic images have been considered and two mixtures (Fig. 2-b) have been generated with a mixing matrix $\mathbf{A} = \begin{bmatrix} 1, & .5 \\ .5, & 1 \end{bmatrix}$. A Gaussian noise have been added to the observations so that the signal to noise ratio (SNR) is approximately equal to 20dB. The resulting estimates obtained with the TDSEP algorithm are presented in Fig. 2-c while those obtained with the proposed approach are presented in Fig. 3 with the three different estimates corresponding to the three presented prior models: the IGM model presents better performances in terms of Performance Index (PI), however it presents lower performances in terms of the chosen distance (correlation coefficient ρ). We point out to the fact that the IGM model needed, for this simulated example, much more iterations (about twice) than those needed

by the other prior models (HMT and CHMF) to get to the desired solution.

The TDSEP algorithm fails to separate properly the original sources since it aims to find, diagonalizing jointly time structure correlation matrices, maximally independent sources. However, the original estimates are not independent as indicated by the inter-correlation coefficient of source 1 with respect to source 2.

On Fig. 4, we present portions of images (64×64 pixel portions) corresponding to the estimates obtained in this simulation example with the prior models, and give as an indication the correlation coefficients of these small portions with respect to the original image. We note that the presented images have been rescaled to be in $[0, 1]$ for presentation purposes. We clearly observe from the figure that the IGM model performs less in terms of edge preservation (as expected from its mathematical description). We also note that the CHMF model outperforms the HMT model (as indicated by the correlation coefficient ρ).

In a second example, two sources (Fig. 5-a) have been considered, both presenting similar geometrical shapes. A mixing matrix of the form $\mathbf{A} = \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}$ have been used to generate the observations with a signal to noise ratio of $\text{SNR} = 15\text{dB}$ (Fig. 5-b). The results obtained with the TDSEP algorithm are represented in figure 5-c: though it succeeds to separate the sources, it presents an unpleasant drawback, it amplifies the noise present in the observations, thus it needs an additional denoising procedure. However it shows that this ICA based algorithm is robust to Gaussian noise (returns a good estimation of the mixing matrix: $\text{PI} = -9.48\text{ dB}$) which is generally a property of ICA based algorithms [4].

The results obtained with the three prior models (IGM, HMT and CHMF) are respectively presented in Fig. 5-{a,b,c}. The IGM model presents similar behavior in this example as in the previous example: it needs much more iterations to reach convergence (in Markov Chain Monte Carlo algorithms, the convergence is a convergence in law). The estimates corresponding to the HMT and CHMF prior models present better performances than the TDSEP algorithm, especially in terms of signal to noise ratio in the final estimates.

A. Joint Separation and Denoising example

In order to highlight the performances of the joint denoising and separation procedure expressed by the BG-prior model (section VI), a third example have been considered (Fig. 7-a) where two observations (Fig. 7-b) have been generated with a square mixing matrix $\mathbf{A} = \begin{bmatrix} 1 & .8 \\ .5 & 1 \end{bmatrix}$ and a relatively low signal to noise ratio ($\text{SNR} = 7\text{dB}$). The results obtained by the TDSEP algorithm are shown on Fig. 7-c: this shows once again the robustness of ICA-algorithms to Gaussian noise, but in the opposite their weakness in estimating the sources and the bothering noise amplification effect.

Simulation results obtained by the prior models (IGM, HMT and CHMF) with the Bernoulli-Gaussian mixture prior are respectively represented on Fig. 8-{a,b} and -c: the IGM model performs well, however it has as the unpleasant effect to smoothen the final estimates (for a comparison on the edge

preservation of each of the three models, see Fig. 4). The HMT model reaches poorer performances (in terms of PI and ρ) while the CHMF reaches better performances in terms of the correlation distance ρ than the two other models, but presents poorer performances than the IGM model in terms of performance index PI.

B. Choice of Wavelet basis

A second set of simulations have been done on the three data sets: but this time the observations have been decomposed on the the ‘‘Daubechies’’ wavelets with 3 vanishing moments (they are highly assymetrical but with minimal support). On Fig. 9, 10 and Fig. 11, we present the box-plots of the obtained estimates with the three prior models (IGM, HTM and CHMF) for the data set 1 (text image), 2 (rice images) and data set 3 (aerial images) respectively compared to the estimation results obtained with ‘‘Symmlet’’ wavelet basis: the estimates we obtain on these data sets do not crucially depend on the choice of the wavelet basis. This is quite expected since the parameters of the prior laws (variances and weights of two Gaussians mixture prior) are estimated within the algorithms and what is really important is rather the sparsity property of the wavelet representation.

C. CHMF prior parameters values β_1 and β_2

We have experimentally observed, on the data sets presented (and others) that, at high signal to noise ratios, the estimates do not significantly change as function of these two parameters as long as $\beta_1 \in [.4, .8] \geq \beta_2 \in [.4, .8]$. We recall that β_1 controls the intra scale correlations while β_2 controls the inter scale correlations in the CHMF model presented in section IV-E. However at lower signal to ratio ratios the estimates depend sensitively on the values they are given: this is expected because the likelihood at low signal to noise rations is less informative and then the prior parameters play a more significant role.

D. HMT-global updating vs. HMT-local updating

In the data sets we considered, we have observed that the obtained estimates in the case of the HMT prior model with its global updating version are very similar to those obtained with HMT prior model with its local updating version, so we present in the simulations only those corresponding to the HTM-global updating version. A discussion and a comparison of the two versions on 1D Markov chain examples is given in [25] without a definite answer on their relative performances. However for the prior transition probabilities (transition probability from a coefficient to its children $\pi_{q'q}$ in equation 25), a transition probability matrix of the form $\begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix}$ has been chosen. Recalling that this transition probabilities are being updated by the likelihood of the data (so we expect that they do play an important role in low signal to noise ratios).

VIII. CONCLUSION

In this paper we addressed the problem of blind source separation of linear and instantaneous noisy mixtures in a Bayesian estimation framework. We addressed the problem in the wavelet domain, that allowed to define appropriate prior models for the wavelet coefficients of the unobserved sources. We considered three models:

- Independent Gaussian Mixture model (IGM), which does not account neither for inter nor intra scale correlations,
- Hidden Markov Tree model (HMT) that accounts for inter scale correlations,
- Contextual Hidden Markov Field model (CHMF) that accounts for both intra and inter-scale correlations.

In our knowledge, even if these models have been used for image denoising or restoration, they have not yet been used in blind source separation, which is the main contribution of this paper.

Simulations have been performed on a set of images and the results were reported and compared to a classical time structure ICA method (the TDSEP algorithm), where we have seen that the wavelet based Bayesian source separation (Bayes-BSS) approach presented in this paper outperforms the ICA-based method for the data sets considered herein. However ICA methods are known to be very fast methods as compared to MCMC sampling based approaches. Approximations of the posterior distributions and suboptimal methods can be considered in order to implement faster solutions by this approach.

A limiting case of the original two Gaussian mixture model was considered, the Bernoulli-Gaussian (BG) mixture model, in order to be able to perform a joint separation and denoising of observations affected by a high noise level. Simulation results that have been reported showed that this approach seems to be promising.

Accounting for inter and intra scale correlations through a Contextual Hidden Markov Field (CHMF) model improves clearly the estimation results of the source images, particularly at low signal to noise ratios, in terms of image discontinuities and edge preservation. However, a work is still to be done on the optimal choice of the Hidden Markov Field constants (β_1, β_2) .

APPENDIX

In this appendix, we will detail the sampling distribution (28), which in this case rewrites:

$$\begin{aligned} p(z_\lambda = \mathbf{q} | \mathbf{w}_x^\lambda, \boldsymbol{\theta}) & \propto \mathcal{N}(\mathbf{w}_x^\lambda | 0, \mathbf{R}_{x|z}) \pi(z_\lambda = \mathbf{q} | \boldsymbol{\theta}_s) \\ & \propto \mathcal{N}(\mathbf{w}_x^\lambda | 0, \mathbf{R}_{x|z}) \prod_i \pi(z_\lambda^{(i)} = q^{(i)} | \boldsymbol{\theta}_s) \end{aligned} \quad (37)$$

where $\mathbf{R}_{x|z}$ is given in section V-A and $q^{(i)} \in \{L, H\}$.

A. Sampling distribution of the IGM model (section IV-C)

For the IGM model, the conditional prior³ $\pi(z_\lambda = q | \boldsymbol{\theta}_s)$ are in fact the prior weights of equation (23):

$$\pi(z_\lambda = q | \boldsymbol{\theta}_s) = p_q \quad (\text{A.1})$$

The conditional sampling distribution of $[p_L, p_H]$ is given by:

$$\begin{aligned} p(p_L, p_H | z_{1, \dots, K_j}) & \propto p(z_{1, \dots, K_j} | p_L, p_H) \pi(p_L, p_H), \\ & \propto \left[\prod_{k_j=1}^{K_j} p(z_{k_j} | p_L, p_H) \right] \mathcal{D}(u_L, u_H), \\ & \propto \mathcal{D}(\gamma_L, \gamma_H), \end{aligned} \quad (\text{A.2})$$

where $\lambda = (j, k_j)$, $\gamma_L = \text{Card}\{z_\lambda = L, \text{ for } k_j = 1, \dots, K_j\}$ and $\mathcal{D}(p_L, p_H | u_L, u_H) \propto p_L^{u_L-1} p_H^{u_H-1}$.

B. Sampling distribution of the HMT variables (section IV-D)

In equation (37), \mathbf{q} is the vector of all the possible combinations of $\mathbf{q} = \{q^{(1)}, \dots, q^{(n)}\} \in \{1, \dots, 2^n\}$ and equation (37) can be rewritten:

$$\begin{aligned} p(\tilde{z}_\lambda = \mathbf{q} | \mathbf{w}_x^\lambda, \boldsymbol{\theta}) & \propto \mathcal{N}(\mathbf{w}_x^\lambda | 0, \mathbf{R}_{x|z}) \pi(\tilde{z}_\lambda = \mathbf{q} | \boldsymbol{\theta}_s) \\ & \propto f_q(w_\lambda) \pi(\tilde{z}_\lambda = \mathbf{q} | \boldsymbol{\theta}_s). \end{aligned}$$

In the following, posterior sampling distributions for $\tilde{z} \in \{1, \dots, 2^n\}$ will be given, the inverse transform to obtain $z_\lambda = [z_\lambda^{(1)}, \dots, z_\lambda^{(n)}]$ is trivial. For the HMT model, two sampling scheme of the hidden variables are possible:

1) *Global updating*: Let $\pi_{qq'} = p(\tilde{z}_{c_\lambda} = q' | \tilde{z}_\lambda = q)$.

For $\lambda = (2, \{1, \dots, K_2\}), \dots, (J, \{1, \dots, K_J\})$, the backward variables $\beta_\lambda(q)$ are iteratively given by:

$$\begin{aligned} \beta_\lambda(q) & = p(T_{C_\lambda} | \tilde{z}_\lambda = q) \\ & = \sum_{\tilde{z}_{c_\lambda}} p(T_{C_\lambda} | \tilde{z}_{c_\lambda}) p(\tilde{z}_{c_\lambda} | \tilde{z}_\lambda = q), \\ & = \sum_{\tilde{z}_{c_\lambda}} p(T_{C_{c_\lambda}} | \tilde{z}_{c_\lambda}) p(w_{c_\lambda} | \tilde{z}_{c_\lambda}) p(\tilde{z}_{c_\lambda} | \tilde{z}_\lambda = q) \end{aligned} \quad (\text{B.1.1})$$

since

$$\begin{aligned} p(T_{C_{c_\lambda}} | \tilde{z}_{c_\lambda}) & = \prod_{r \in C_\lambda} p(T_{C_r} | \tilde{z}_r = q_r) = \prod_{r \in C_\lambda} \beta_r(q_r), \\ p(w_{c_\lambda} | \tilde{z}_{c_\lambda}) & = \prod_{r \in C_\lambda} p(w_r | \tilde{z}_r = q_r) = \prod_{r \in C_\lambda} f_{q_r}(w_r), \\ p(\tilde{z}_{c_\lambda} | \tilde{z}_\lambda = q) & = \prod_{r \in C_\lambda} p(\tilde{z}_r = q_r | \tilde{z}_\lambda = q) = \prod_{r \in C_\lambda} \pi_{qq_r}, \end{aligned}$$

then

$$\beta_\lambda(q) = \prod_{r \in C_\lambda} \left(\sum_{q_r} \beta_r(q_r) f_{q_r}(w_r) \pi_{qq_r} \right). \quad (\text{B.1.2})$$

$$\beta_{(1, \dots)}(q) = 1, \forall q \in \{1, \dots, 2^n\}.$$

³The index (i) is dropped to alleviate the notations.

For $\lambda = (J, \{1, \dots, K_J\})$, the *forward* variables $\alpha_\lambda(q)$ at the lowest resolution are given by:

$$\alpha_\lambda(q) = p(w_\lambda, \tilde{z}_\lambda = q) = f_q(w_\lambda) p_q. \quad (\text{B.1.3})$$

and the marginal probabilities $\gamma_\lambda(q)$ are then given by:

$$\gamma_\lambda(q) = p(\tilde{z}_\lambda = q | T_{(J, K_J)}) = \frac{\alpha_\lambda(q) \beta_\lambda(q)}{\sum_q \alpha_\lambda(q) \beta_\lambda(q)}. \quad (\text{B.1.4})$$

We have then, iteratively, for $\lambda = (J - 1, \{1, \dots, K_{(J-1)}\})$, \dots , $(1, \{1, \dots, K_1\})$:

$$\begin{aligned} p(\tilde{z}_\lambda = q | \tilde{z}_{J, \dots, j+1}, T_{\setminus \lambda}) \\ \propto f_q(w_\lambda) p(T_{C_\lambda} | \tilde{z}_\lambda) \pi(\tilde{z}_\lambda = q | \tilde{z}_{P_\lambda} = q'), \\ \propto f_q(w_\lambda) \beta_\lambda(q) \pi_{q'q}. \end{aligned} \quad (\text{B.1.5})$$

2) *Local updating*: For $\lambda = (J, \{1, \dots, K_J\})$:

$$\begin{aligned} p(\tilde{z}_\lambda = q | T_{(J, K_J)}, \tilde{z}_{\Lambda \setminus \lambda}) \\ \propto p(T_{(J, K_J)} | \tilde{z}_\Lambda) p(\tilde{z}_{\Lambda \setminus \lambda} | \tilde{z}_\lambda = q) \pi(\tilde{z}_\lambda = q), \\ \propto f_q(w_\lambda) p(\tilde{z}_{C_\lambda} | \tilde{z}_\lambda = q) p_q, \\ \propto f_q(w_\lambda) \left[\prod_{r \in C_\lambda} p(\tilde{z}_r = q_r | \tilde{z}_\lambda = q) \right] p_q \\ \propto f_q(w_\lambda) \left[\prod_{r \in C_\lambda} \pi_{q_r q} \right] p_q. \end{aligned} \quad (\text{B.2.1})$$

For $\lambda = (J - 1, \{1, \dots, K_{J-1}\})$, \dots , $(2, \{1, \dots, K_2\})$:

$$\begin{aligned} p(\tilde{z}_\lambda = q | T_{(J, K_J)}, \tilde{z}_{\Lambda \setminus \lambda}) \\ \propto p(T_{(J, K_J)} | \tilde{z}_\Lambda) p(\tilde{z}_{\Lambda \setminus \lambda} | \tilde{z}_\lambda = q), \\ \propto p(w_\lambda | \tilde{z}_\lambda = q) p(\tilde{z}_\lambda = q | \tilde{z}_{P_\lambda} = q') p(\tilde{z}_{C_\lambda} | \tilde{z}_\lambda = q), \\ \propto f_q(w_\lambda) p(\tilde{z}_\lambda = q | \tilde{z}_{P_\lambda} = q') \\ \times \left[\prod_{r \in C_\lambda} p(\tilde{z}_r = q_r | \tilde{z}_\lambda = q) \right], \\ \propto f_q(w_\lambda) \pi_{q'q} \left[\prod_{r \in C_\lambda} \pi_{q_r q} \right]. \end{aligned} \quad (\text{B.2.2})$$

For $\lambda = (1, \{1, \dots, K_1\})$:

$$\begin{aligned} p(\tilde{z}_\lambda = q | T_{(J, K_J)}, \tilde{z}_{\Lambda \setminus \lambda}) \\ \propto p(T_{(J, K_J)} | \tilde{z}_\Lambda) p(\tilde{z}_{\Lambda \setminus \lambda} | \tilde{z}_\lambda = q), \\ \propto p(w_\lambda | \tilde{z}_\lambda = q) p(\tilde{z}_\lambda = q | \tilde{z}_{P_\lambda} = q'), \\ \propto f_q(w_\lambda) \pi_{q'q}. \end{aligned} \quad (\text{B.2.3})$$

REFERENCES

- [1] Pierre Comon, "Independent Component Analysis, a new concept?," *Signal processing, Special issue on Higher-Order Statistics*, Elsevier, vol. 36, no. 3, pp. 287–314, April 1994.
- [2] Dinh-Tuan Pham and Philippe Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. on Signal Processing*, vol. 45, no. 7, pp. 1712–1725, July 1997.
- [3] Kevin Knuth, "A Bayesian approach to source separation," in *Proceedings of Independent Component Analysis Workshop*, 1999, pp. 283–288.
- [4] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, *Independent Component Analysis*, John Wiley, New York, 2001.
- [5] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle, "Independent component analysis based on higher-order statistics only," *IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, 1996, pp. 356–359.
- [6] Jianting Cao, Noboru Murata, Shun ichi Amari, Andrzej Cichocki, and Tsunehiro Takeda, "A robust approach to independent component analysis of signals with high-level noise measurements," *IEEE Trans. on Neural Networks*, vol. 14, no. 3, pp. 631–645, May 2003.
- [7] Adel Belouchrani, Karim Abed-Meraim, and Jean-François Cardoso, "A blind source separation technique using second-order statistics," *IEEE Trans. on Signal Processing*, vol. 45, no. 2, pp. 434–444, February 1997.
- [8] Lutz Mogedey and Heinz Georg Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.
- [9] Andreas Ziehe and Klaus-Robert Müller, "TDSEP - an efficient algorithm for blind source separation using time structure," in *International Conference on Artificial Neural Networks*, September 1998, pp. 675–680.
- [10] Klaus-Robert Müller, Petra Philips, and Andreas Ziehe, "JADE_{TD}: Combining higher-order statistics and temporal information for blind source separation (with noise)," in *International Workshop on Independent Component Analysis and Blind Source Separation*, 1999.
- [11] Jean François Cardoso, "Higher-order contrasts for independent component analysis," in *Neural Computation*, vol. 11 of *MIT Letters*, pp. 157–192, 1999.
- [12] Daniel B. Rowe, *Correlated Bayesian Factor analysis*, Ph.D. thesis, Department of Statistics, Univ. of California, Riverside, 1998.
- [13] Ali Mohammad-Djafari, "A Bayesian approach to source separation," in *Bayesian Inference and Maximum Entropy Methods, AIP*, Boise (IH), USA, July 1999, MaxEnt, pp. 221–244, AIP.
- [14] Harri Lappalainen, "Ensemble learning for independent component analysis," in *First International Workshop on Independent Component Analysis*, 1999, pp. 7–12.
- [15] Hichem Snoussi, *Bayesian approach to source separation. Applications in imagery*, Ph.D. thesis, University of Paris-Sud, Orsay, France, September 2003.
- [16] Rizwan A. Choudrey and Stephen J. Roberts, "Bayesian ICA with hidden Markov model sources," in *ICA*, Nara, Japan, April 2003.
- [17] Adel Belouchrani and Moeness G. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Trans. on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, November 1998.
- [18] Michael Zibulevsky and Barak A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computations*, vol. 13, no. 4, pp. 863–882, 2001.
- [19] Mahieddine M. Ichir and Ali Mohammad-Djafari, "Séparation de sources modélisées par des ondelettes," in *GRETSI*, Paris, France, September 2003.
- [20] Matthew S. Crouse, Robert D. Nowak, and Richard G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. on Signal Processing*, vol. 46, no. 4, April 1998.
- [21] Eero P. Simoncelli and Edward H. Adelson, "Noise removal via a bayesian wavelet coring," *ICIP*, vol. 1, pp. 379–382, September 1996.
- [22] Pierre Moulin and Juan Liu, "Analysis of multiresolution image denoising schemes using generalized-Gaussian and complexity priors," *IEEE Trans. on Information Theory*, vol. 45, pp. 909–919, April 1999.
- [23] Anestis Antoniadis, David Leporini, and Jean-Christophe Pesquet, "Wavelet thresholding for some classes of non-gaussian noise," *Statistica Neerlandica*, vol. 56, no. 4, pp. 434–453, 2002.
- [24] Jean-Christophe Pesquet, Hamid Krim, David Leporini, and E. Hamman, "A Bayesian approach to best basis selection," *Atlanta (GA), USA, 1996, ICASSP*, vol. 5, pp. 2634–2637.
- [25] Christian P. Robert, Gilles Celeux, and Jean Diebolt, "Bayesian estimation of hidden Markov models: a stochastic implementation," *Statistics and Probability Letters*, vol. 16, pp. 77–83, 1993.
- [26] Aapo Hyvärinen, "New approximations of differential entropy for independent component analysis and projection pursuit," in *Advances in Neural Information Processing Systems 10*, 1998, MIT Press, pp. 273–279.
- [27] Christian Robert, *L'analyse statistique bayésienne*, Economica, Paris, France, 1992.
- [28] Mahieddine M. Ichir and Ali Mohammad-Djafari, "Bayesian Wavelet Based Statistical Signal and Image Separation," in *Bayesian Inference and Maximum Entropy Methods*, Jackson Hole (WY), USA, August 2003, MaxEnt, pp. 417–428, AIP.

to appear on the IEEE Transactions on Image Processing

C. Sampling distribution of the CHMF variables (section IV-E)

For $\lambda = (1, \{1, \dots, K_1\})$:

$$\begin{aligned} p(\tilde{z}_\lambda = q | T_{(J, K_J)}, \tilde{z}_{\nu_\lambda}, \tilde{z}_{\mathcal{P}_\lambda}) &\propto p(w_\lambda | \tilde{z}_\lambda = q) p(\tilde{z}_\lambda = q | \tilde{z}_{\nu_\lambda}) p(\tilde{z}_{\mathcal{P}_\lambda} = q' | \tilde{z}_\lambda = q), \\ &\propto f_q(w_\lambda) \exp\left(\beta_1 \sum_{r \in \nu_\lambda} \delta(\tilde{z}_r = q) + \beta_2 \delta(\tilde{z}_{\mathcal{P}_\lambda} = q)\right) \end{aligned} \quad (\text{C.1})$$

For $\lambda = (J - 1, \cdot), \dots, (2, \cdot)$:

$$\begin{aligned} p(\tilde{z}_\lambda = q | T_{(J, K_J)}, \tilde{z}_{\nu_\lambda}, \tilde{z}_{\mathcal{P}_\lambda}, \tilde{z}_{\mathcal{C}_\lambda}) &\propto p(w_\lambda | \tilde{z}_\lambda = q) p(\tilde{z}_\lambda = q | \tilde{z}_{\nu_\lambda}, \tilde{z}_{\mathcal{C}_\lambda}) p(\tilde{z}_{\mathcal{P}_\lambda} = q' | \tilde{z}_\lambda = q), \\ &\propto f_q(w_\lambda) \exp\left(\beta_1 \sum_{r \in \nu_\lambda} \delta(\tilde{z}_r = q) + \beta_2 \left(\sum_{r \in \mathcal{C}_\lambda} \delta(\tilde{z}_r = q) + \delta(\tilde{z}_{\mathcal{P}_\lambda} = q)\right)\right) \end{aligned} \quad (\text{C.2})$$

For $\lambda = (J, \{1, \dots, K_J\})$:

$$\begin{aligned} p(\tilde{z}_\lambda = q | T_{(J, K_J)}, \tilde{z}_{\nu_\lambda}, \tilde{z}_{\mathcal{C}_\lambda}) &\propto p(w_\lambda | \tilde{z}_\lambda = q) p(\tilde{z}_\lambda = q | \tilde{z}_{\nu_\lambda}, \tilde{z}_{\mathcal{C}_\lambda}), \\ &\propto f_q(w_\lambda) \exp\left(\beta_1 \sum_{r \in \nu_\lambda} \delta(\tilde{z}_r = q) + \beta_2 \sum_{r \in \mathcal{C}_\lambda} \delta(\tilde{z}_r = q)\right) \end{aligned} \quad (\text{C.3})$$

- [29] James W. Miskin, *Ensemble Learning for Independent Component Analysis*, Ph.D. thesis, Selwyn College, Cambridge University, 2000.
- [30] Rizwan A. Choudrey, *Variational Methods for Independent Component Analysis*, Ph.D. thesis, Univ. of Oxford, 2002.
- [31] Dinh-Tuan Pham and Jean-François Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, September 2001.
- [32] H. Snoussi, G. Patanchon, J.F. Macías-Pérez, A. Mohammad-Djafari, and J. Delabrouille, "Bayesian blind component separation for cosmic microwave background observations," in *Bayesian Inference and Maximum Entropy Methods*. MaxEnt, August 2001, pp. 125–140, AIP.
- [33] Elie L. Benroya, *Séparation de plusieurs sources sonores avec un seul capteur*, Ph.D. thesis, Univ. de Rennes I, 2003.
- [34] Hichem Snoussi and Ali Mohammad-Djafari, "Fast joint separation and segmentation of mixed images," *Journal of Electronic Imaging*, vol. 13, no. 2, pp. 349–361, April 2004.
- [35] Alfred Haar, *Zur Theorie der orthogonalen Funktionensysteme*, Ph.D. thesis, Göttingen, Germany, 1909.
- [36] Stephane G. Mallat, *a Wavelet Tour of Signal Processing*, Academic Press, 1999.
- [37] Ingrid Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
- [38] Stephane G. Mallat, "A theory of multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, July 1989.
- [39] David S. Taubman and Michael W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards, and Practice*, Kluwer Academic, 2001.
- [40] David L. Donoho and Iain M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
- [41] David L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995.
- [42] Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," February 1989, vol. 77 of *Proceedings of the IEEE*, pp. 257–286.
- [43] Odile Macchi and Eric Moreau, "Adaptive unsupervised separation of discrete sources," in *Signal Processing*, 1999, vol. 73, pp. 49–66.
- [44] Eric Moulines, Jean-François Cardoso, and Elisabeth Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *ICASSP*, Munich, Germany, April 1997.
- [45] Aapo Hyvärinen, "Independent component analysis in the presence of gaussian noise by maximizing joint likelihood," *Neurocomputing*, vol. 22, pp. 49–67, 1998.

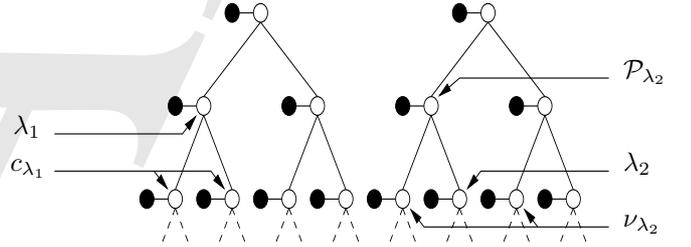


Fig. 1. Binary tree representation of the wavelet coefficients for 1D signal wavelet decomposition: ● Wavelet coefficients, ○ Associated hidden variables (corresponding to a quad tree for 2D signals).

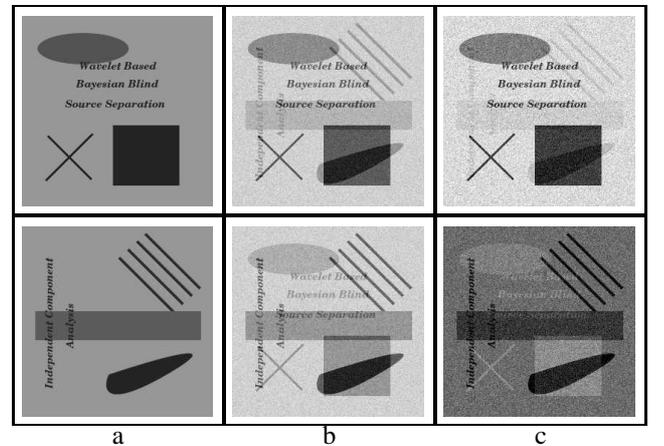


Fig. 2. a) original sources: $\rho_{(s,s)}[1,2] = 0.2225$, b) mixture images: $\rho_{(x,x)}[1,2] = 0.8271$, $\mathbf{A} = [[1, .5]; [.5, 1]]$ and SNR = 20dB, c) estimated sources with the TDSEP algorithm with 4 time lag correlation matrices: $PI_{dB} = -9.0437$ and $\rho_{(s,s)}[1,2] = -0.1381$

	X_1	X_2	TDSEP		IGM		HGMT _{gu}		HGMF	
$\rho_{(s,s)}[1,1]$	0.9163	0.6286	0.9447	-0.1365	0.9835	0.2331	0.9888	0.2629	0.9900	0.2611
$\rho_{(s,s)}[1,2]$	0.5341	0.8625	0.3413	0.7985	0.2296	0.9728	0.2507	0.9835	0.2579	0.9854
$\rho_{(s,s)}[2,1]$	0.8077	0.6840	0.5812	0.1477	0.9663	0.0638	0.9642	-0.0431	0.9675	-0.0124
$\rho_{(s,s)}[2,2]$	0.4964	0.6707	-0.1302	0.6633	-0.0102	0.9352	0.0342	0.9468	0.0716	0.9553
$\rho_{(s,s)}[1,1]$	0.6213	0.3072	0.5045	-0.0901	0.8510	-0.0503	0.7930	0.0799	0.8314	0.0017
$\rho_{(s,s)}[1,2]$	0.3816	0.5973	0.0437	0.4046	-0.1301	0.9481	-0.3615	0.9253	-0.2273	0.9350

TABLE I

TABLE SUMMARIZING THE PERFORMANCES FOR THE THREE SIMULATION EXAMPLES PRESENTED IN FIG. 3 (TOP ROW), FIG. 6 (MIDDLE ROW) AND FIG. 8 (BOTTOM ROW)

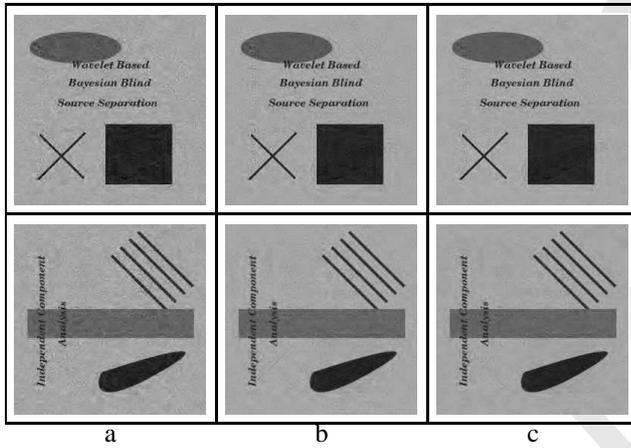


Fig. 3. estimated sources corresponding to the: a) IGM model: $PI_{dB} = -52.3816$ and $\rho_{(s,s)} = 0.2224$, b) HMT (-global updating) model: $PI_{dB} = -29.1059$ and $\rho_{(s,s)}[1,2] = 0.2813$, c) CHMF ($\beta_1 = \beta_2 = .5$) model: $PI_{dB} = -27.0361$ and $\rho_{(s,s)}[1,2] = 0.2877$

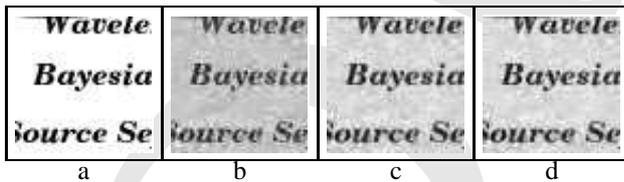


Fig. 4. a) enlargement of (gray rescaled)-regions (64×64 pixels) of the: a) source image of Fig. 2-a (top), b) IGM-estimated image of Fig. 3-a (top): $\rho_{(s,s)}[1,1] = 0.9451$, c) HMT-estimated image of Fig. 3-b (top): $\rho_{(s,s)}[1,1] = 0.9685$, d) CHMF-estimated image of Fig. 3-c (top): $\rho_{(s,s)}[1,1] = 0.9708$.

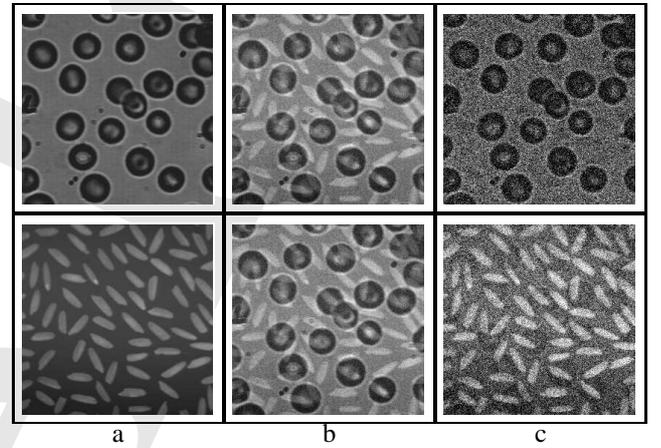


Fig. 5. a) original sources: $\rho_{(s,s)}[1,2] = -0.0278$, b) mixture images: $\rho_{(x,x)}[1,2] = 0.9108$, $\mathbf{A} = \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}$ and SNR = 15dB, c) estimated sources with the TDSEP algorithm with 4 time lag correlation matrices: $PI_{dB} = -9.4850$ and $\rho_{(s,s)}[1,2] = -0.5631$

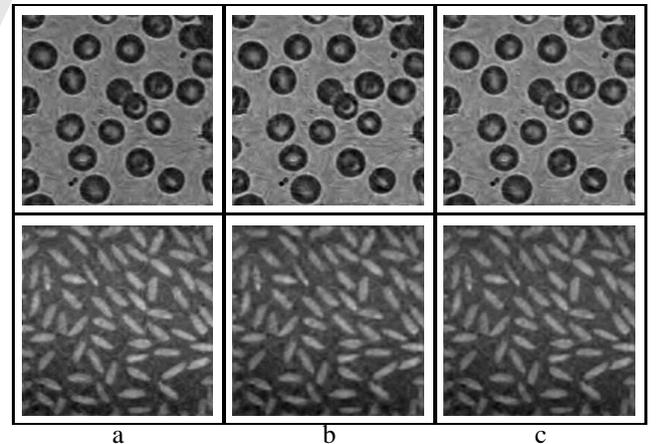


Fig. 6. estimated sources corresponding to the: a) IGM model: $PI_{dB} = -22.4201$ and $\rho_{(s,s)} = 0.0165$, b) HMT (-global updating) model: $PI_{dB} = -22.4109$ and $\rho_{(s,s)}[1,2] = -0.0555$, c) CHMF ($\beta_1 = \beta_2 = .5$) model: $PI_{dB} = -25.2147$ and $\rho_{(s,s)}[1,2] = -0.0255$

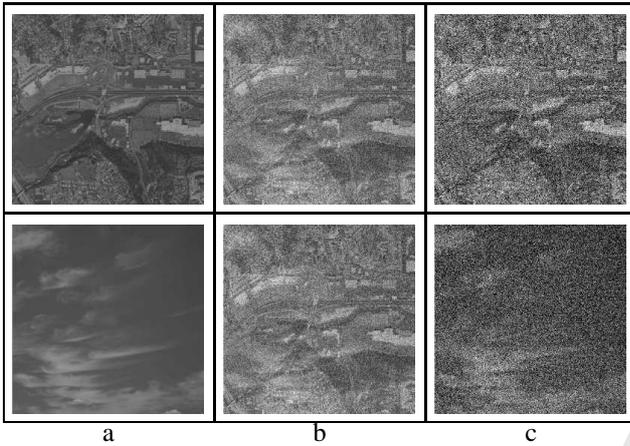


Fig. 7. a) original sources: $\rho_{(s,s)}[1, 2] = -0.1117$, b) mixture images: $\rho_{(x,x)}[1, 2] = 0.4806$, $\mathbf{A} = \begin{bmatrix} 1 & .8 \\ -.5 & 1 \end{bmatrix}$ and $\text{SNR} = 7\text{dB}$, c) estimated sources with the TDSEP algorithm with 4 time lag correlation matrices: $PI_{\text{dB}} = -11.4810$ and $\rho_{(\hat{s},\hat{s})}[1, 2] = -0.6955$

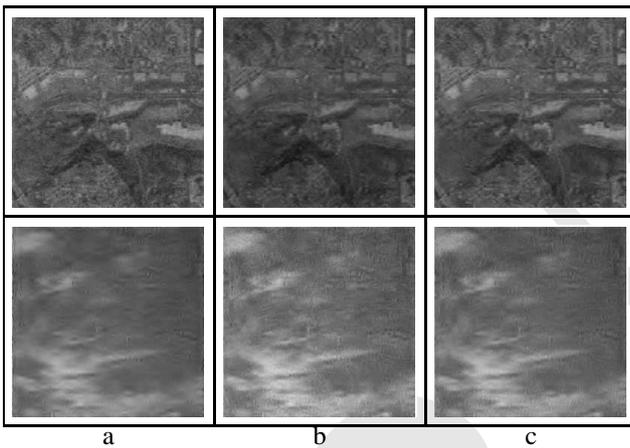


Fig. 8. estimated sources corresponding to the: a) IGM model: $PI_{\text{dB}} = -21.9871$ and $\rho_{(\hat{s},\hat{s})} = -0.1412$, b) HMT (-global updating) model: $PI_{\text{dB}} = -9.2129$ and $\rho_{(\hat{s},\hat{s})}[1, 2] = -0.2625$, c) CHMF ($\beta_1 = .7, \beta_2 = .6$) model: $PI_{\text{dB}} = -15.2715$ and $\rho_{(\hat{s},\hat{s})}[1, 2] = -0.2011$

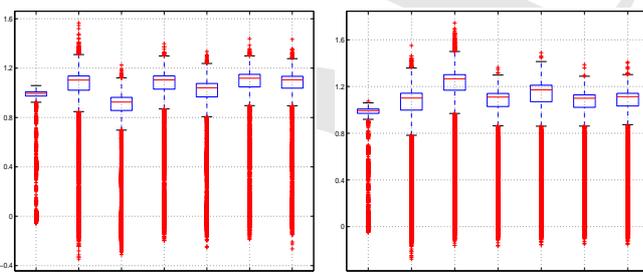


Fig. 9. box-plots of source 1 (left) and source 2 (right) for data set of Fig. 2 where columns 1 to 7 represent respectively 1) original source; 2) and 3) IGM-based estimates with ‘Symmlets-6’ wavelets (column 2) and with ‘Daubechies-3’ wavelets (column 3); 4) and 5) HMT-based estimates with ‘Symmlets-6’ and ‘Daubechies-3’ respectively; 6) and 7) CHMF-based estimates with ‘Symmlets-6’ and ‘Daubechies-3’ respectively

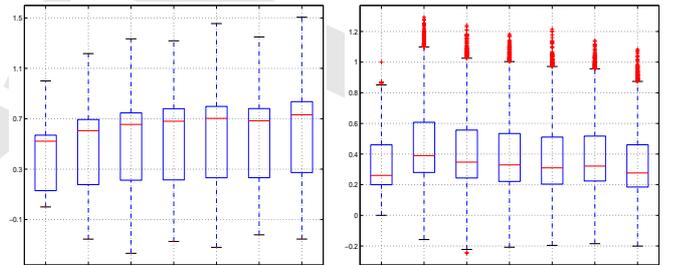


Fig. 10. box-plots of source 1 (left) and source 2 (right) for data set of Fig. 5 where columns 1 to 7 represent respectively 1) original source; 2) and 3) IGM-based estimates with ‘Symmlets-6’ wavelets (column 2) and with ‘Daubechies-3’ wavelets (column 3); 4) and 5) HMT-based estimates with ‘Symmlets-6’ and ‘Daubechies-3’ respectively; 6) and 7) CHMF-based estimates with ‘Symmlets-6’ and ‘Daubechies-3’ respectively

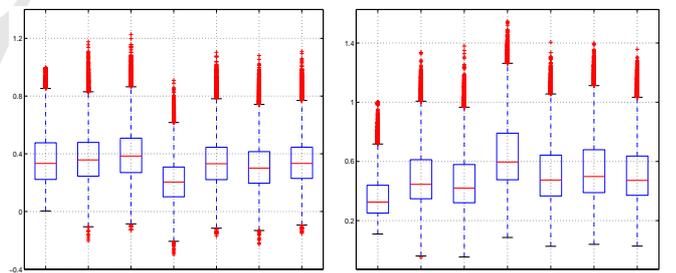


Fig. 11. box-plots of source 1 (left) and source 2 (right) for data set of Fig. 7 where columns 1 to 7 represent respectively 1) original source; 2) and 3) IGM-based estimates with ‘Symmlets-6’ wavelets (column 2) and with ‘Daubechies-3’ wavelets (column 3); 4) and 5) HMT-based estimates with ‘Symmlets-6’ and ‘Daubechies-3’ respectively; 6) and 7) CHMF-based estimates with ‘Symmlets-6’ and ‘Daubechies-3’ respectively