A Bayesian Approach for Blind Separation of Sparse Sources

Cédric Févotte and Simon J. Godsill, *Member, IEEE*

***Abstract*—We present a Bayesian approach for blind separation of linear instantaneous mixtures of sources having a sparse representation in a given basis. The distributions of the coefficients of the sources in the basis are modeled by a Student t distribution, which can be expressed as a scale mixture of Gaussians, and a Gibbs sampler is derived to estimate the sources, the mixing matrix, the input noise variance and also the hyperparameters of the Student t distributions. The method allows for separation of underdetermined (more sources than sensors) noisy mixtures. Results are presented with audio signals using a modified discrete cosine transform basis and compared with a finite mixture of Gaussians prior approach. These results show the improved sound quality obtained with the Student t prior and the better robustness to mixing matrices close to singularity of the Markov chain Monte Carlo approach.**

***Index Terms*—Bayesian estimation, blind source separation (BSS), independent component analysis, Markov chain Monte Carlo (MCMC) methods, sparse representations.**

I. Introduction

BLIND source separation (BSS) consists in estimating n signals (the sources) from the sole observation of m mixtures of them (the observations). In this paper, we consider linear instantaneous mixtures of time series: at each time index, the observations are a linear combination of the sources at the same time index.

(Over)determined ($m \geq n$) nonnoisy linear instantaneous mixtures have been widely studied, within the field of independent component analysis (ICA), assuming independently and identically distributed (i.i.d.) sources and using higher order statistics (see [1], [2] for a survey), or using correlation (e.g., [3]), nonstationarity (e.g., [4]), or both (e.g., [5]), leading to second order statistics based methods. These methods require at least mutual decorrelation of the sources (strict independence being required by high order statistics based methods) and little prior information on each source (typically, the sign of the kurtosis or mild hypotheses on spectra). These methods have proved to perform well, and one can reasonably argue that there is little left to be done in separation of (over)determined nonnoisy linear instantaneous mixtures.

Difficulties arise when dealing with mixtures which are possibly underdetermined and possibly noisy. The underdetermined case in particular is very challenging because contrary to (over)determined mixtures, estimating the mixing system is not sufficient for reconstruction of the sources, since, for $m < n$, the mixing matrix is not invertible. Then, it appears that separation of underdetermined mixtures requires important prior information on the sources to allow their reconstruction. Prior information can also be helpful for reconstructing the sources in noisy environments.

In this paper, we propose to tackle the general linear instantaneous model (possibly underdetermined, possibly noisy) using the assumption of *sparsity* of the sources on a given basis. This assumption means that only a few coefficients of the decomposition of the sources on the basis are significantly nonzero. The use of sparsity to handle the general linear instantaneous model, has arisen in several papers in the areas of learning [6]–[8] and source separation [9]–[13]. In the source separation context, source time series can be assumed to be sparse in time (speech sequences with silences are considered in [9], which describes a source separation method adapted from the learning method in [7]) but, more generally, to have a sparse representation on a given dictionary, possibly overcomplete (number of elements of the dictionary greater than the length of the signals). The aim of these methods then becomes the estimation of the coefficients of the sources in the dictionary and not the time series themselves. The time series are then reconstructed from the estimated coefficients. Besides the methods in [12] and [13], which rely directly on disjoint time-frequency supports of the sources and propose deterministic methods, the methods in [6]–[11] model the distributions of the coefficients of the sources (the *priors*) with a distribution gathering most of its probability around zero and presenting heavy tails, thus modeling sparsity. The methods then proceed in two steps: the first one consists in *learning* the mixing matrix from the data, the second one in *inferring* the sources from the learnt mixing matrix and the data, with respect to the chosen prior on the sources.[1]

In [7]–[10], the coefficients of the sources are given a Laplacian prior. Learning is performed via clustering in the mixtures space in [10] and via a maximum likelihood (ML) approach in [7]–[9]. The latter approach requires marginalization of the sources in the likelihood expression, which is done via a Laplacian approximation of the integral in [7] and [9] and via a variational approximation of the Laplacian prior in [8]. Inferring is realized by linear programming in [7], [9], and [10] while obtained as a byproduct of the EM optimization procedure used in [8].

Manuscript received January 28, 2005; revised July 16, 2005. This work was supported by the European Commission funded Research Training Network HASSIP under HPRN-CT-2002-00285. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shoji Makino.

The authors are with the Signal Processing Group, Engineering Department, Cambridge University, Cambridge CB2 1PZ, U.K. (e-mail: cf269@eng.cam.ac.uk; sjg@eng.cam.ac.uk).

Digital Object Identifier 10.1109/TSA.2005.858523

[1] In an iterative fashion, the learning step often requires inferring some statistics of the sources conditionally upon the data and the current estimate of the mixing matrix and possibly other parameters.

In [6] and [11], the coefficients of the sources are given a finite mixture of Gaussians (FMoG) prior distribution with two or three states [one state "off" corresponding to a Gaussian with very small variance, possibly zero, and high probability, one or two state(s) "on" corresponding to Gaussians with big variance and small probability]. In [6], learning is achieved via a full ML optimization, which is easily done taking advantage of the source coefficients being Gaussian conditionally upon the states. Inferring the sources requires inferring the states, which is done by a Gibbs sampling step. A pseudo-ML criterion is used for learning in [11] by truncating the full ML criterion to its $n+1$ dominant terms, which is equivalent to the assumption that only one source can be "on" at each time-frequency point. Inferring of the sources is then done by marginalization over all the possible states.

The authors of [11] present results with speech signals decomposed on a modified discrete cosine transform (MDCT) orthonormal basis [14]. The method proves to perform as well as the full ML criterion approach (with decreased computational burden) and better than [7] in the presence of noise. The use of a basis provides equivalence between representations in the time domain and transform domain, and separation can be simply performed in the transform domain instead of the time domain. The use of an overcomplete dictionary is appealing because it allows potentially sparser representations, but is less straightforward than the use of a basis because of the one-to-many mapping between the time and transform domains [10].

Motivated by the results of Student t modeling for sparse Bayesian learning [15], [16], we consider a Student t prior on the source coefficients, which leads to sparse modeling when the degrees of freedom parameter is low. The work in this paper can be interpreted as separation of possibly noisy and underdetermined linear instantaneous mixtures of mutually independent Student t distributed sources, or in an other parlance, possibly overcomplete and noisy ICA with sparse Student t prior. Preliminary results were presented in [17]. We will use the following key property of the Student t: It can be expressed as a scale mixture of Gaussians (SMoG), whose scale function corresponds to the conjugate prior on the variances of the Gaussians (see Section II-B2). Within a Bayesian approach, this property allows us to derive a Gibbs sampler, a standard Markov chain Monte Carlo (MCMC) simulation method, to sample from the posterior conditional distribution of the parameters upon the data, which include not only the sources and the mixing matrix but also *hyperparameters* of the Student t priors. In particular, we propose a method to sample from the degrees of freedom of the Student t priors (which are not constrained to be the identical for each source). In the context of sparse representations the degrees of freedom parameter has an important role since it measures the sharpness of the Student t distribution and, thus, the degree of sparsity of the sequence of coefficients of a source. The resulting samples of all the parameters then allow for computation of any point estimate (as well as interval estimates), for example minimum mean-square error (MMSE) estimates.

Out of the context of sparse representations, pionneering work on Bayesian approaches to source separation is described in [18]–[22], where gradient based methods are designed for various priors on the sources. Other interesting references are [23] and [24] where a Gibbs sampler and an iterated conditional modes method are described to perform separation of mutually *correlated* sources, assumed conditionally Gaussian upon a covariance matrix which is given an inverted-Wishart (conjugate) prior. However, audio examples in previous work on sparse sources separation tended to show that the assumption of mutual independence of the coefficients of decomposition of the sources on a chosen dictionary is reasonable (this is discussed in [10]). This assumption of mutual independence furthermore allows us to assume different degrees of freedom for each sequence of source coefficients. MCMC methods were also applied to separation of discrete sources in [25] and [26] and to separation of autoregressive sources in [27].

The paper is organized as follows. Section II introduces notations and assumptions. In Section III, we present briefly the Gibbs sampler, which requires the posterior distributions of each parameter conditional upon the data and the other parameters, evaluated in Section IV. Section V presents separation results with determined and mostly underdetermined mixtures of audio signals decomposed on a MDCT basis. Our approach is compared with the one in [11] and we show improved audio quality resulting from the Student t model with respect to the FMoG model. The MCMC approach is also shown to be more robust to mixing matrix initialization and mixing matrices which are close to singularity. Conclusions and perspectives are given in Section VI.

II. Model and Assumptions

A. Model and Aim

We consider the standard linear instantaneous model where the observations at time t are noisy linear combinations of the sources at same time t and the signals are of finite length N, such that $\forall t = 1, \ldots, N$

$$\mathbf{x}_t = \mathbf{A}\,\mathbf{s}_t + \mathbf{n}_t \tag{1}$$

where $\mathbf{x}_t = [x_{1,t}, \ldots, x_{m,t}]^T$ is a vector of size m containing the observations, $\mathbf{s}_t = [s_{1,t}, \ldots, s_{n,t}]^T$ is a vector of size n containing the sources and $\mathbf{n}_t = [n_{1,t}, \ldots, n_{m,t}]^T$ is a vector of size m containing additive noise. Variables without time index t denote whole sequences of samples, e.g., $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and $x_1 = [x_{1,1}, \ldots, x_{1,N}]$.

The aim of the following work is to estimate the sources $\mathbf{s}$ and the mixing matrix $\mathbf{A}$ up to the standard BSS indeterminacy on gain and order, that is, compute $\hat{\mathbf{s}}$ and $\hat{\mathbf{A}}$ such that (ideally)

$$\hat{\mathbf{A}} = \mathbf{A}\,\mathbf{D}\,\mathbf{P} \tag{2}$$

$$\hat{\mathbf{s}} = \mathbf{P}^T\,\mathbf{D}^{-1}\,\mathbf{s} \tag{3}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a permutation matrix.

B. Assumptions

In this section, we describe the assumptions made on the sources and the noise.

1) Time Domain/Transform Domain: We assume that we are given a basis on which the sources adopt a sparse representation. Again, this means that only a low proportion of coefficients of the decompositions are significantly different from zero. Let $\mathbf{\Phi}$

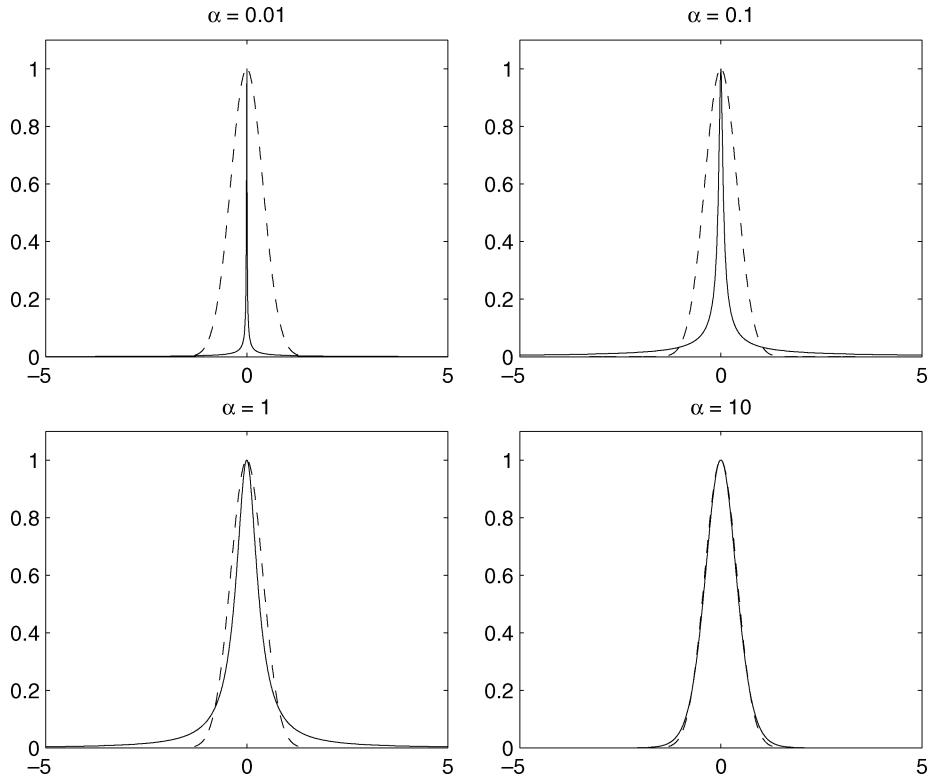

Fig. 1. Student t densities for $\alpha \in \{0.01, 0.1, 1, 10\}$ with equal value at the mode . The dash-lined plot is the Gaussian density with variance $1/2\,\pi$.

be a $N \times N$ invertible matrix defining such a basis. We denote by $\tilde{y} = y\,\mathbf{\Phi}$ the decomposition of a time series y in $\mathbf{\Phi}$. The decomposition of the observations is written

$$\tilde{\mathbf{x}} = \mathbf{x}\,\mathbf{\Phi} \tag{4}$$

$$= \mathbf{A}\,\tilde{\mathbf{s}} + \tilde{\mathbf{n}} \tag{5}$$

or, equivalently, $\forall k = 1, \ldots, N$

$$\tilde{\mathbf{x}}_k = \mathbf{A}\,\tilde{\mathbf{s}}_k + \tilde{\mathbf{n}}_k. \tag{6}$$

k indexes the coefficients of decomposition of the time signals in the basis. Because $\mathbf{\Phi}$ is a basis, (6) is strictly equivalent to (1), which means that separation can be performed equivalently either in the time domain or in the transform domain. In the following, we work in the transform domain.

2) Model of Sparsity: We assume that the sequences of coefficients $\tilde{s}_i$ are i.i.d. with Student t distribution $t(\alpha_i, \lambda_i)$ defined as

$$p(\tilde{s}_{i,k} \mid \alpha_i, \lambda_i) = \frac{\Gamma(\frac{\alpha_i+1}{2})}{\lambda_i \sqrt{\alpha_i \pi}\,\Gamma(\frac{\alpha_i}{2})} \left(1 + \frac{1}{\alpha_i}\left(\frac{\tilde{s}_{i,k}}{\lambda_i}\right)^2\right)^{-(\alpha_i+1)/2}. \tag{7}$$

α_i is the degrees of freedom and λ_i is a scale parameter. With $\lambda_i = 1$ and $\alpha_i = 1$, the Student t distribution is equal to the standard Cauchy distribution, and it tends to the standard Gaussian distribution as α_i goes to infinity. Fig. 1 plots Student t densities for several values of α_i, with equal mode, i.e., setting $\lambda_i = \Gamma((\alpha_i+1)/2)/\sqrt{\alpha_i\,\pi}\,\Gamma(\alpha_i/2)$ for each density. Fig. 1 shows that for small α_i, the Student t density gathers most of its probability mass around zero and exhibits "fatter tails" than the normal distribution. The Student t distribution is, thus, a relevant model for sparsity.

Note that the i.i.d. assumption of each sequence $\tilde{s}_i$ is only a *working assumption*.[2] Indeed, this assumption does not mean that our method is bound to fail when the coefficient sequences are not i.i.d. (which is likely to happen for instance with MDCT coefficients of audio signals, because of the existence of time and frequency structures). It only means that we choose not to use the inner correlation of the sequences, though this information could be used to design future separation schemes (see [15] and discussion in Section VI).

The Student t distribution can be expressed as a SMoG [28], such that

$$p(\tilde{s}_{i,k} \mid \alpha_i, \lambda_i) = \int_0^{+\infty} \mathcal{N}(\tilde{s}_{i,k} \mid 0, v_{i,k})\,\mathcal{IG}\left(v_{i,k} \mid \frac{\alpha_i}{2}, \frac{2}{\alpha_i \lambda_i^2}\right) dv_{i,k} \tag{8}$$

where $\mathcal{N}(x \mid \mu, v)$ and $\mathcal{IG}(x \mid \gamma, \beta)$ are the Gaussian and inverted-Gamma distributions, defined in Appendix I. Thus, introducing the auxiliary random variable $v_{i,k}$, $p(\tilde{s}_{i,k} \mid \alpha_i, \lambda_i)$ can be interpreted as a marginal density of the joint distribution $p(\tilde{s}_{i,k}, v_{i,k} \mid \alpha_i, \lambda_i)$, defined by

$$p(\tilde{s}_{i,k}, v_{i,k} \mid \alpha_i, \lambda_i) = p(\tilde{s}_{i,k} \mid v_{i,k})\, p(v_{i,k} \mid \alpha_i, \lambda_i) \tag{9}$$

with

$$p(\tilde{s}_{i,k} \mid v_{i,k}) = \mathcal{N}(\tilde{s}_{i,k} \mid 0, v_{i,k}) \quad \text{and}$$
$$p(v_{i,k} \mid \alpha_i, \lambda_i) = \mathcal{IG}\left(v_{i,k} \mid \frac{\alpha_i}{2}, \frac{2}{\alpha_i\,\lambda_i^2}\right). \tag{10}$$

[2]This expression is used in several papers of Pham (see, e.g., [4]).

The fact that $\tilde{s}_{i,k}$ is Gaussian conditionally upon the variance $v_{i,k}$ of the Gaussian distribution in (8) is a very convenient property. Furthermore, the distribution of the variances $v_{i,k}$ is inverted-Gamma, which corresponds to the conjugate prior of $v_{i,k}$ in the expression of $\mathcal{N}(s_{i,k} \mid 0, v_{i,k})$, which means that the variance $v_{i,k}$ conditionally upon $\tilde{s}_{i,k}$, α_i and λ_i will also be inverted-Gamma and, thus, easy to sample from (see Section IV-D).

The Student t can, thus, be interpreted as an infinite sum of Gaussians, which contrasts with the finite sums of Gaussians used in [6] and [11]. The Laplacian prior used in [7] and [9] can also be expressed as a SMoG, with an exponential density on the variance $v_{i,k}$ [28]. Unfortunately, this density does not lead to a distribution of $v_{i,k}$ conditionally upon $\tilde{s}_{i,k}$ (and the scale parameter) straightforward to sample from. In comparison with the Laplacian prior, the Student t prior has the advantage to offer a supplementary hyperparameter α_i which controls the sharpness of the distribution.

In the following, we define $\mathbf{v}_k = [v_{1,k}, \ldots, v_{n,k}]^T$, $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]$ and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]$.

3) Mutual Independence: We assume that the coefficients sequences of the sources are mutually independent, such that $p(\tilde{\mathbf{s}}) = \prod_{i=1}^{n} p(\tilde{s}_i)$. As pointed out in [10], the assumption of mutual independence of the *coefficients* of the sources on the basis may be considered more realistic than the mutual independence of the sources themselves, which is the standard assumption of ICA methods.

4) Noise Properties: We assume that $\tilde{\mathbf{n}}$ is a i.i.d. Gaussian noise with covariance $\sigma^2 \mathbf{I}_m$ and with σ unknown. We point out that when an orthonormal basis is used (i.e., $\boldsymbol{\Phi}^{-1} = \boldsymbol{\Phi}^T$), $\mathbf{n}$ is equivalently a Gaussian i.i.d. noise with covariance $\sigma^2 \mathbf{I}_m$. Nevertheless, if one assumes the expected squared deviation of the signal known, then the principle of maximum entropy states that the Gaussian is the least biased probability assignment that is consistent with this knowledge.

We now present a MCMC approach to estimate the set of parameters of interest $\{\tilde{\mathbf{s}}, \mathbf{A}, \sigma\}$ together with $\mathbf{v}$ and the hyperparameters $\{\boldsymbol{\alpha}, \boldsymbol{\lambda}\}$. We define $\boldsymbol{\theta} = \{\tilde{\mathbf{s}}, \mathbf{A}, \sigma, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\lambda}\}$, and $\boldsymbol{\theta}_{-\mathbf{y}}$ will denote the set of parameters in $\boldsymbol{\theta}$ except $\mathbf{y}$. For instance, $\boldsymbol{\theta}_{-\mathbf{A}} = \{\tilde{\mathbf{s}}, \sigma, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\lambda}\}$.

III. Presentation of the Gibbs Sampler

We derive in the following a Gibbs sampler to generate samples from the *posterior* distribution $p(\boldsymbol{\theta} \mid \tilde{\mathbf{x}})$. The obtained samples $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}\}$ then allow for computation of any required point estimate (as well as interval estimates), a standard one being the MMSE estimate defined by

$$\hat{\boldsymbol{\theta}}_{\mathrm{MMSE}} = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \tilde{\mathbf{x}})\, d\boldsymbol{\theta} \tag{11}$$

and which can be approximated by

$$\hat{\boldsymbol{\theta}}_{\mathrm{MMSE}} \approx \frac{1}{K} \sum_{l=1}^{K} \boldsymbol{\theta}^{(l)}. \tag{12}$$

The Gibbs sampler only requires to be able to sample from the posterior distribution of certain subsets of parameters conditional upon data $\tilde{\mathbf{x}}$ and the remaining parameters [29], [30]. Let $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ denote a partition of $\boldsymbol{\theta}$. The Gibbs sampler works as follows:

Initialize $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_M^{(0)}\}$
for $k = 1 : K + K_{\mathrm{Burnin}}$ **do**
$\boldsymbol{\theta}_1^{(k)} \sim p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(k-1)}, \ldots, \boldsymbol{\theta}_M^{(k-1)}, \tilde{\mathbf{x}})$
$\boldsymbol{\theta}_2^{(k)} \sim p(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(k)}, \boldsymbol{\theta}_3^{(k-1)}, \ldots, \boldsymbol{\theta}_M^{(k-1)}, \tilde{\mathbf{x}})$
$\boldsymbol{\theta}_3^{(k)} \sim p(\boldsymbol{\theta}_3 \mid \boldsymbol{\theta}_1^{(k)}, \boldsymbol{\theta}_2^{(k)}, \boldsymbol{\theta}_4^{(k-1)} \ldots, \boldsymbol{\theta}_M^{(k-1)}, \tilde{\mathbf{x}})$
$\vdots$
$\boldsymbol{\theta}_M^{(k)} \sim p(\boldsymbol{\theta}_M \mid \boldsymbol{\theta}_1^{(k)}, \boldsymbol{\theta}_2^{(k)}, \ldots, \boldsymbol{\theta}_{M-1}^{(k)}, \tilde{\mathbf{x}})$
end for.

K_{Burnin} represents the number of iterations required before the Markov chain $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots\}$ reaches its stationary distribution $p(\boldsymbol{\theta} \mid \tilde{\mathbf{x}})$. Thereafter, all samples are drawn from the desired stationary distribution. MCMC methods have the advantage to generate samples from the whole support of $p(\boldsymbol{\theta} \mid \tilde{\mathbf{x}})$ and, thus, to give an overall panorama of the *posterior* distribution of the parameters. When looking for a point estimate of $\boldsymbol{\theta}$, the MCMC approach prevents from falling into local modes of the *posterior* distribution, which is a common drawback of standard optimization methods, like expectation maximization or gradient type methods, which target point estimates (such as maximum *a posteriori* estimates) directly.

To implement the Gibbs sampler for $\boldsymbol{\theta} = \{\tilde{\mathbf{s}}, \mathbf{A}, \sigma, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\lambda}\}$, we now need to derive the various densities of each parameter conditional upon the others and the data.

IV. Conditional Densities

A. General Outline

The conditional distribution $p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{-k}, \tilde{\mathbf{x}})$ of each parameter conditional upon the other parameters and the data is written

$$p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{-k}, \tilde{\mathbf{x}}) = \frac{p(\boldsymbol{\theta} \mid \tilde{\mathbf{x}})}{p(\boldsymbol{\theta}_{-k} \mid \tilde{\mathbf{x}})}. \tag{13}$$

Bayes' theorem gives [31]

$$p(\boldsymbol{\theta} \mid \tilde{\mathbf{x}}) = \frac{p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\tilde{\mathbf{x}})} \tag{14}$$

and it follows:

$$p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{-k}, \tilde{\mathbf{x}}) = \frac{p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}_{-k} \mid \tilde{\mathbf{x}})\, p(\tilde{\mathbf{x}})} \tag{15}$$

$$\propto p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}). \tag{16}$$

Thus, the required conditional distributions are proportional to the *likelihood* of the data times the *priors* on the parameters.

1) Prior Structure: Assuming independence of the mixing matrix, the noise variance and the sources parameters, we have

$$p(\boldsymbol{\theta}) = p(\tilde{\mathbf{s}}, \mathbf{A}, \sigma, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) \tag{17}$$

$$= p(\tilde{\mathbf{s}} \mid \mathbf{v})\, p(\mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\lambda})\, p(\boldsymbol{\alpha})\, p(\boldsymbol{\lambda})\, p(\mathbf{A})\, p(\sigma). \tag{18}$$

Assuming no specific prior information about $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$, $\mathbf{A}$, and σ, these parameters will be given noninformative prior distributions. Uniform priors will be used for $\boldsymbol{\alpha}$ and $\mathbf{A}$ while Jeffrey's prior $p(x) = 1/x$ will be favored for $\boldsymbol{\lambda}$ and σ to keep their role of

"pivotal quantities" (scale parameters) in their posterior distribution [32]. However, since the size of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$, $\mathbf{A}$, and σ do not increase with N (contrary to $\mathbf{s}$ and $\mathbf{v}$), the influence of their prior distribution fades out when reasonable amount of data is available.

2) Likelihood: Under Gaussian noise and sources coefficients i.i.d. assumptions, the likelihood is written

$$p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta}) = p(\tilde{\mathbf{x}} \mid \mathbf{A}, \tilde{\mathbf{s}}, \sigma) \tag{19}$$

$$= \prod_{k=1}^{N} \mathcal{N}(\tilde{\mathbf{x}}_k \mid \mathbf{A}\,\tilde{\mathbf{s}}_k, \sigma^2 \mathbf{I}_m) \tag{20}$$

$$= \frac{1}{(2\pi\sigma^2)^{Nm/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^{N} \|\tilde{\mathbf{x}}_k - \mathbf{A}\tilde{\mathbf{s}}_k\|_F^2\right) \tag{21}$$

where $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian distribution defined in Appendix I. Using the prior structure and the expression of the likelihood, we now give resulting conditional densities of each parameter upon the others and the data. In the following sections, we summarize resulting distributions and sampling schemes; further details are given in the appendices.

B. Sampling $\tilde{\mathbf{s}}$

We show in Appendix II-A

$$p(\tilde{\mathbf{s}} \mid \boldsymbol{\theta}_{-\tilde{\mathbf{s}}}, \tilde{\mathbf{x}}) = \prod_{k=1}^{N} \mathcal{N}(\tilde{\mathbf{s}}_k \mid \boldsymbol{\mu}_{\tilde{\mathbf{s}}_k}, \boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_k}) \tag{22}$$

where $\boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_k} = \left((1/\sigma^2)\,\mathbf{A}^T\mathbf{A} + \mathrm{diag}\,(\mathbf{v}_k)^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\tilde{\mathbf{s}}_k} = (1/\sigma^2)\,\boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_k}\,\mathbf{A}^T\,\tilde{\mathbf{x}}_k$ [and where $\mathrm{diag}\,(\mathbf{u})$ is the diagonal matrix whose main diagonal is given by $\mathbf{u}$]. Note that $\tilde{\mathbf{s}}_k$ being Gaussian conditionally upon $\mathbf{v}_k$, the latter expressions are nothing but the formula of linear estimation of a Gaussian vector parameter in Gaussian noise [31].

C. Sampling $\mathbf{A}$ and σ

It is possible to sample $\mathbf{A}$ and σ separately, by evaluating and sampling from $p(\mathbf{A} \mid \boldsymbol{\theta}_{-\mathbf{A}}, \tilde{\mathbf{x}})$ and $p(\sigma \mid \boldsymbol{\theta}_{-\sigma}, \tilde{\mathbf{x}})$, as done in [17], but it is also possible to sample directly from $p(\mathbf{A}, \sigma \mid \boldsymbol{\theta}_{-(\mathbf{A},\sigma)}, \tilde{\mathbf{x}})$, which is a better strategy as it is recommended to sample as many parameters as possible together to accelerate convergence of the Gibbs sampler [33], [34]. Sampling $(\mathbf{A}^{(k)}, \sigma^{(k)})$ from $p(\mathbf{A}, \sigma \mid \boldsymbol{\theta}_{-(\mathbf{A},\sigma)}, \tilde{\mathbf{x}})$ is equivalent to sampling $\sigma^{(k)}$ from $p(\sigma \mid \boldsymbol{\theta}_{-(\mathbf{A},\sigma)}, \tilde{\mathbf{x}})$ and then sampling $\mathbf{A}^{(k)}$ from $p(\mathbf{A} \mid \boldsymbol{\theta}_{-(\mathbf{A},\sigma)}, \sigma^{(k)}, \tilde{\mathbf{x}})$ [30].

It is shown in Appendix II-C that with uninformative prior $p(\sigma) = 1/\sigma$, we have

$$(\sigma^2 \mid \boldsymbol{\theta}_{-(\mathbf{A},\sigma)}, \tilde{\mathbf{x}}) \sim \mathcal{IG}(\alpha_\sigma, \beta_\sigma) \tag{23}$$

with

$$\alpha_\sigma = \frac{(N-n)\,m}{2} \quad \text{and}$$

$$\beta_\sigma = \frac{2}{\sum_{i=1}^{m}\sum_{k=1}^{N} \tilde{x}_{i,k}^2 - \left(\sum_{k=1}^{N} \tilde{x}_{i,k}\tilde{\mathbf{s}}_k^T\right)\left(\sum_{k=1}^{N} \tilde{\mathbf{s}}_k\tilde{\mathbf{s}}_k^T\right)^{-1}\left(\sum_{k=1}^{N} \tilde{x}_{i,k}\tilde{\mathbf{s}}_k\right)}. \tag{24}$$

Now, we need to sample from $p(\mathbf{A} \mid \boldsymbol{\theta}_{-\mathbf{A}}, \tilde{\mathbf{x}})$. Let $\mathbf{r}_1, \ldots, \mathbf{r}_m$ be the $n \times 1$ vectors denoting the transposed rows of $\mathbf{A}$, such that $\mathbf{A}^T = [\mathbf{r}_1 \ \ldots \ \mathbf{r}_m]$. We show in Appendix II-B that with uninformative uniform prior $p(\mathbf{A}) \propto 1$, we have

$$p(\mathbf{A} \mid \boldsymbol{\theta}_{-\mathbf{A}}, \tilde{\mathbf{x}}) = \prod_{i=1}^{m} \mathcal{N}(\mathbf{r}_i \mid \boldsymbol{\mu}_{\mathbf{r}_i}, \boldsymbol{\Sigma}_{\mathbf{r}}) \tag{25}$$

with $\boldsymbol{\Sigma}_{\mathbf{r}} = \sigma^2\,(\sum_{k=1}^{N} \tilde{\mathbf{s}}_k\,\tilde{\mathbf{s}}_k^T)^{-1}$ and $\boldsymbol{\mu}_{\mathbf{r}_i} = (1/\sigma^2)\,\boldsymbol{\Sigma}_{\mathbf{r}} \sum_{k=1}^{N} \tilde{x}_{i,k}\,\tilde{\mathbf{s}}_k$. Note the posterior mutual independence of the rows.

Because of the inherent source separation indeterminacy on gain, a scale factor has to be fixed, whether on the mixing matrix or on the sources. We discuss several possibilities.

- *Indeterminacy removed by constraints on* $\mathbf{A}$: One option is to set a row $\mathbf{r}_{i'}$ to fixed values, for instance ones. We then sample from the other rows and the scale parameters λ_i. This has the advantage to initialize easily $\mathbf{A}$ when a rough estimate of the mixing matrix is available (for instance provided by a clustering method in the space of the observations [10], [35]): Only divide the columns of $\mathbf{A}$ by the values of the chosen row $\mathbf{r}_{i'}$, and initialize the Gibbs sampler with the resulting matrix. However, this can be problematic if one source has a low contribution in the corresponding sensor i'. Another option is to set the norms of the columns of $\mathbf{A}$ to a fixed value, but implies sampling the columns of $\mathbf{A}$ on a sphere, which is more complex as it induces posterior dependence between the rows.
- *Indeterminacy removed by constraining the sources*: We can set the scale parameters λ_i to fixed values and then sample from all the rows of the mixing matrix. However, we found out in practice that the initialization of $\mathbf{A}$ is difficult. Initializing the mixing matrix with a matrix of zeros does not give satisfactory results because a column of $\mathbf{A}$ gets stuck to nearly zero values, and initializing $\mathbf{A}$ with entries in a correct range of values is not straightforward. The ideal situation is when a rough estimate of $\mathbf{A}$ is available. In that case, the scale parameters of the Student t distributions can be estimated with a standard method [36] (by setting the degrees of freedom to a reasonable fixed value or possibly estimating them) providing values which can be used as initializations of λ_i in the Gibbs sampler. Then we can proceed by sampling from all the rows of $\mathbf{A}$ in order to refine the mixing matrix estimate.

To obtain the results presented in Section V, we chose for its simplicity the option consisting in setting the first row of $\mathbf{A}$ to ones and sampling from the other rows and the scale parameters of the sources. However, we acknowledge that the other schemes are to be investigated, but readily fit into our sampling framework.

D. Sampling $\mathbf{v}$

Since the likelihood does not depend on the parameters $\{\mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\lambda}\}$, their posterior distributions are conditionally independent of $\tilde{\mathbf{x}}$. The variances $v_{i,k}$ having an inverted-Gamma (conjugate) prior, we easily obtain (Appendix II-D)

$$p(\mathbf{v} \mid \boldsymbol{\theta}_{-\mathbf{v}}, \tilde{\mathbf{x}}) = \prod_{k=1}^{N}\prod_{i=1}^{n} \mathcal{IG}\left(v_{i,k} \mid \gamma_{v_i}, \beta_{v_{i,k}}\right) \tag{26}$$

with $\gamma_{v_i} = (\alpha_i + 1)/2$ and $\beta_{v_{i,k}} = 2/(\tilde{s}_{i,k}^2 + \alpha_i\,\lambda_i^2)$.

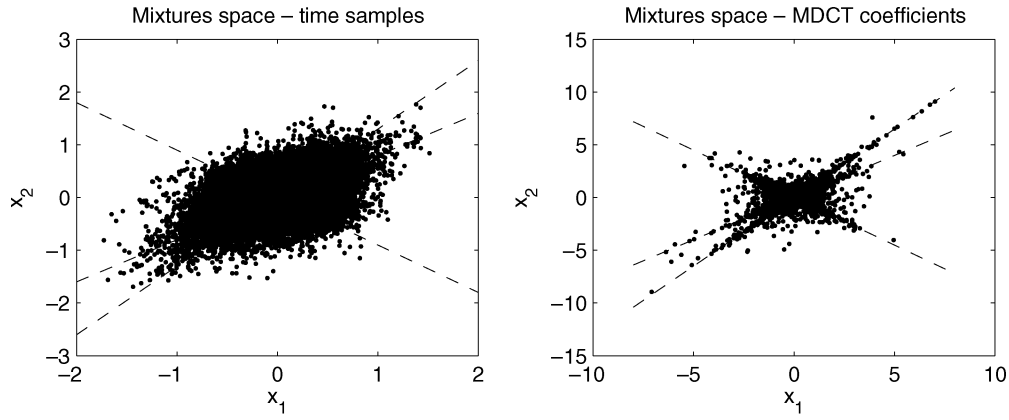

Fig. 2. (a) Scatter plot of x_1 w.r.t x_2. (b) Scatter plot of $\bar{x}_1$ w.r.t $\bar{x}_2$; the dashed lines represent the directions of the mixing matrix.

E. Sampling $\boldsymbol{\alpha}$

We have (Appendix II-E)

$$p(\boldsymbol{\alpha} \mid \boldsymbol{\theta}_{-\boldsymbol{\alpha}}, \tilde{\mathbf{x}}) \propto \prod_{i=1}^{n} \frac{P_i^{-(\alpha_i/2+1)}}{\Gamma(\frac{\alpha_i}{2})^N} \left(\frac{\alpha_i \lambda_i^2}{2}\right)^{\alpha_i N/2} \exp\left(-\frac{\alpha_i \lambda_i^2}{2} R_i\right) p(\alpha_i) \quad (27)$$

with $R_i = \sum_{k=1}^{N} 1/v_{i,k}$ and $P_i = \prod_{k=1}^{N} v_{i,k}$. We choose an uninformative uniform prior on α_i and set $p(\alpha_i) \propto 1$. As the distribution of $(\boldsymbol{\alpha} \mid \boldsymbol{\theta}_{-\boldsymbol{\alpha}}, \tilde{\mathbf{x}})$ is not straightforward to sample from, and since the precise value α_i for each source is unlikely to be important provided it is within an appropriate small range, in practice, we sample $\boldsymbol{\alpha}$ from a uniform grid of discrete values with probability masses proportional to (27). Note that Metropolis–Hastings sampling can also be adopted for this task without the need to grid values.

F. Sampling $\boldsymbol{\lambda}$

Finally, with the uninformative prior $p(\lambda_i) = 1/\lambda_i$, we have (Appendix II-E)

$$(\lambda_i^2 \mid \boldsymbol{\theta}_{-\boldsymbol{\lambda}}, \tilde{\mathbf{x}}) \sim \mathcal{G}(\gamma_{\lambda_i}, \beta_{\lambda_i}) \quad (28)$$

with $\gamma_{\lambda_i} = (\alpha_i N)/2$ and $\beta_{\lambda_i} = 2/(\alpha_i R_i)$, and where $\mathcal{G}(\gamma, \beta)$ is the Gamma distribution given in Appendix I.

G. Summary

At this point, all the posterior conditional distributions required for the Gibbs sampler have been presented. All we have to do now is decompose the observations $\mathbf{x}$ into the given basis $\boldsymbol{\Phi}$, and apply the Gibbs sampler described in Section III to the sequences of coefficients $\tilde{\mathbf{x}}$. The several steps of the sampler are recapitulated in Table I and emphasizes the dependencies between the several parameters. Samples from $\tilde{\mathbf{s}}$ and $\mathbf{A}$ are, thus, obtained, and MMSE estimates are computed using (12). Estimates of the original sources $\mathbf{s}$ can be obtained by applying $\boldsymbol{\Phi}^{-1}$ to the estimate of $\tilde{\mathbf{s}}$. We stress that thanks to the Student t model, expressed as a SMoG with inverted-Gamma distributed variances, all the posterior conditional distributions are easy to sample from. In the following section, we consider the separation of audio signals decomposed on a MDCT basis, a local cosine lapped transform [14] which has proven to give good sparse approximations of audio signals, with many coding applications [37], [38]. We first consider a noisy underdetermined mixing in Section V-A and then a noisy determined mixing in Section V-B. Robustness of the methods over several mixing matrices are investigated in Section V-C. The results are then commented in Section V-D, where we discuss separation quality, robustness to the mixing matrix and convergence properties of the methods, as well as computational issues.

TABLE I
GIBBS SAMPLER FOR SOURCE SEPARATION OF STUDENT t SOURCES

Initialize $\boldsymbol{\theta}^{(0)} = \{\tilde{\mathbf{s}}^{(0)}, \mathbf{A}^{(0)}, \sigma^{(0)}, \mathbf{v}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\lambda}^{(0)}\}$
for $k = 1 : K + K_{Burnin}$ **do**
  $\tilde{\mathbf{s}}^{(k)} \sim p(\tilde{\mathbf{s}} | \mathbf{A}^{(k-1)}, \sigma^{(k-1)}, \mathbf{v}^{(k-1)}, \tilde{\mathbf{x}})$
  $\sigma^{(k)} \sim p(\sigma | \tilde{\mathbf{s}}^{(k)}, \tilde{\mathbf{x}})$
  $\mathbf{A}^{(k)} \sim p(\mathbf{A} | \tilde{\mathbf{s}}^{(k)}, \sigma^{(k)}, \tilde{\mathbf{x}})$
  $\mathbf{v}^{(k)} \sim p(\mathbf{v} | \tilde{\mathbf{s}}^{(k)}, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)})$
  $\boldsymbol{\alpha}^{(k)} \sim p(\boldsymbol{\alpha} | \mathbf{v}^{(k)}, \boldsymbol{\lambda}^{(k-1)})$
  $\boldsymbol{\lambda}^{(k)} \sim p(\boldsymbol{\lambda} | \mathbf{v}^{(k)}, \boldsymbol{\alpha}^{(k)})$
end for

V. RESULTS

A. Underdetermined Mixture

We first study a mixture of $n = 3$ audio sources (speech, piano, guitar) with $m = 2$ observations. The mixing matrix is given in Table II. We set $\sigma = 0.03$, which corresponds to approximatively 20-dB noise on each observation. The signals are sampled at 8 kHz with length $N = 65\,356$ (≈ 8 s). We used a MDCT basis to decompose the observations, using a sine bell and 50% overlap, yielding a time resolution of 64 ms (half the window length). Time resolutions of 32 and 128 ms were also tried but led to slightly poorer results. Fig. 2 shows scatter plots of the observations in the time domain and the transform domain. Fig. 2(b) shows the *sparsification* induced by the MDCT decomposition: The directions of the mixing matrix columns clearly appear, contrary to Fig. 2(a).

We ran 5000 iterations of the Gibbs sampler, and MMSE estimates of $\mathbf{A}$ and $\tilde{\mathbf{s}}$ where computed from the final 1000 samples. The different parameters were initialized with the values

TABLE II
ESTIMATES OF $\mathbf{A}$ FOR THE UNDERDETERMINED MIXTURE

Original matrix
$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 0.8 & 1.3 & -0.9 \end{bmatrix}$
t + MCMC
$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ 0.7827 & 1.2950 & -0.9021 \\ (\pm 0.0026) & (\pm 0.0018) & (\pm 0.0028) \end{bmatrix}$
FMoG + EM
$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ 0.7545 & 1.3249 & -0.9001 \end{bmatrix}$

shown in the tabular at the bottom of the page (using MATLAB notation).

The discrete values from which the degrees of freedom α_i are sampled from are chosen linearly spaced between 0.05 and 5, with step size 0.05.

We compared our method (referred to as "t + MCMC") with the one in [11], which uses a mixture of two Gaussians to model the source coefficients distribution: one state "on" corresponding to a high variance Gaussian, one state "off" corresponding to a small variance Gaussian (in fact, the method intrinsically sets the "off" state variance to zero to solve identifiability problems inherent to the model; see [11]). The method, referred to as "FMoG + EM," estimates the mixing matrix, the noise covariance and the weights in the mixtures of Gaussians with a EM-based procedure. MMSE estimates of the sources are then inferred from the estimated parameters and the data. The method was run for 50 iterations, and convergence was observed after 30 iterations.

Fig. 3 plots samples drawn from the second row $\mathbf{r}_2$ of $\mathbf{A}$, $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$, and σ throughout the iterations of the Gibbs sampler. Table II shows the mixing matrices estimated by the two methods, together with the standard variations in the case of the t + MCMC method (the FMoG + EM method only provides point estimates).

Fig. 4 plots the distributions of the coefficients of the *original* sources compared to the estimated Student t densities of each source.

Table III provides separation quality criteria for the sources estimates given by the two methods. The criteria used are described in [39]. Basically, the source-to-distortion ratio (SDR) provides an overall separation performance criterion, the source-to-interferences ratio (SIR) measures the level of interferences from the other sources in each source estimate, source-to-noise ratio (SNR) measures the error due to the additive noise on the sensors, and the source-to-artifacts ratio (SAR) measures the level of artifacts in the source estimates. The higher the ratios, the better the quality of estimation. We point out that the performance criteria are invariant to a change of basis, so that figures can be computed either on the time sequences ($\hat{\mathbf{s}}$ compared to $\mathbf{s}$) or the MDCT coefficients ($\hat{\tilde{\mathbf{s}}}$ compared to $\tilde{\mathbf{s}}$). The estimated sources can be listened to at http://www-sigproc.eng.cam.ac.uk/~cf269/ieee_sap05/sound_files.html, which is, perhaps, the best way to assess the audio quality of the results.

B. Determined Mixture

We now give results on a determined mixture ($m = n$). The same sources are used, with mixing matrix given in Table IV. We set $\sigma = 0.1$, which approximatively corresponds to 9.5-dB noise on each observation. As before MDCT decompositions were performed using a time resolution of 64 ms and 5000 iterations of the Gibbs sampler were run (with same initializations), MMSE estimates being computed from the last 1000 iterations. The FMoG + EM method was used with 30 iterations, and convergence was observed after 20 iterations. For comparison, we also show the results provided by the standard ICA algorithm JADE [40] applied to $\tilde{\mathbf{x}}$, which estimates a separating matrix via joint-diagonalization of a set of cumulant matrices, and apply the obtained matrix to the data, without denoising of the sources estimates.

Fig. 5 plots samples drawn from $\mathbf{r}_2$, $\mathbf{r}_3$, $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$, and σ throughout the iterations of the Gibbs sampler. Table IV shows the estimated mixing matrices, Table V shows quality criteria of the estimated sources.

C. Robustness to Mixing Matrix

In this section, we study the robustness of the two methods over several mixing matrices, in particular, over matrices with "close" columns. We used the same audio sources as before, though we now consider only 1s excerpts of them. We designed four mixing matrices of the following form:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ \tan\psi_1 & \tan\psi_2 & \tan\psi_3 \end{bmatrix}. \tag{29}$$

We considered the four sets of values of $\{\psi_1, \psi_2, \psi_3\}$ indicated in Table VI. For visibility, we plot the independent component directions generated by the several matrices in Fig. 6. The first case corresponds to widely spread directions, the second and third cases correspond to two closely spread directions with a third remote direction, and the fourth case corresponds to three closely spread directions. The initialization of the parameters in the Gibbs sampler was chosen as before; the mixing matrix in the FMoG + EM method was initialized with a matrix of the form of (29) with nearly zero angles (the method does not accept strictly zero angles); however, random initializations lead

$\tilde{\mathbf{s}}$	$\mathbf{r}_2$	σ	$\mathbf{v}$	$\boldsymbol{\alpha}$	$\boldsymbol{\lambda}$
$\texttt{ones}(n, N)$	$[0 \quad 0 \quad 0]$	0.1	$\texttt{ones}(n, N)$	$\texttt{ones}(1, n)$	$0.1 * \texttt{ones}(1, n)$

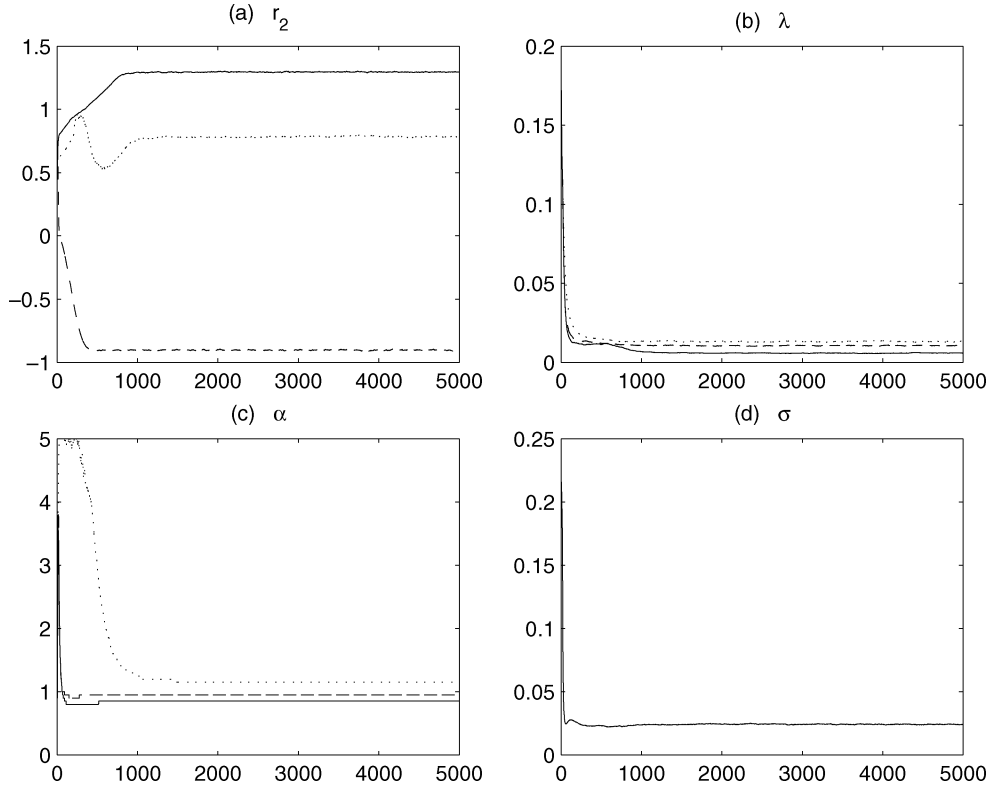

Fig. 3. Samples from the different parameters obtained with the Gibbs sampler for the underdetermined mixture.

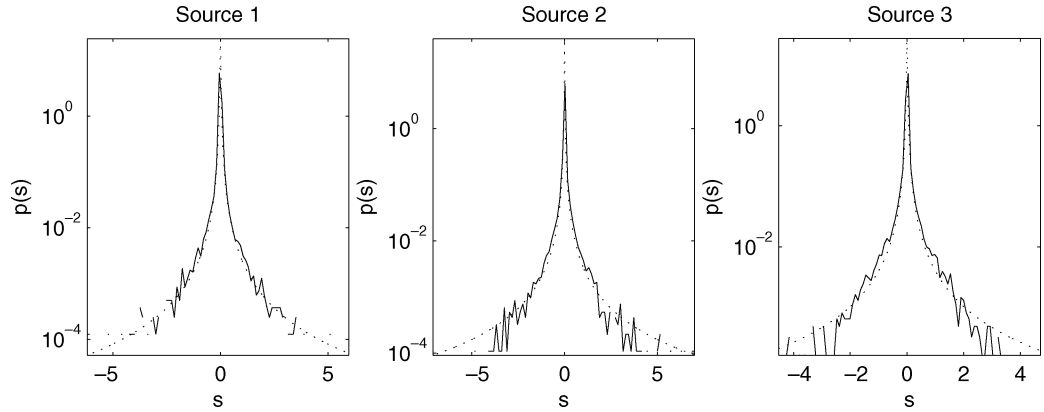

Fig. 4. Histograms of coefficients $\bar{s}_i$ (full lines) compared to the estimated Student t densities (dot lines) . Note the Y axis log cale.

to similar results. Noise with power $\sigma = 0.03$ was added on the observations, providing input SNRs in the range 20–30 dB.

The estimates of $\{\psi_1, \psi_2, \psi_3\}$ provided by both methods are given in Table VI. Corresponding separation quality criteria are given in Table VII.

D. Discussion

1) Separation Quality Assessment: A major advantage of the Student t model over the FMoG model is sound quality. These can be perceived by listening to the sound samples, but also by looking at the SARs of the source estimates which are higher using the t model in all the presented simulations. The t model provides estimates which sound much more natural that the FMoG model. Fig. 4 shows the very good fit between the estimated Student t distributions and the histograms of the MDCT coefficients of the sources.

The main advantage of the FMoG model is noise reduction as measured with SNR. SNRs of the sources estimates are better with the FMoG model than with the t model. We think this is because the model sets the "off" state variance to zero and, thus, performs a thresholding of the MDCT coefficients, which might eliminate small coefficients originated from the additive noise on the observations.

The t model gives slightly poorer interference rejection in the underdetermined case and similar performances to the FMoG model in the determined case (and no interference can be heard over 30-dB SIR). The better SIRs obtained for both methods in the determined case are explained by the fact that in that context

TABLE III
PERFORMANCE CRITERIA FOR THE UNDERDETERMINED MIXTURE

	$\hat{s}_1$				$\hat{s}_2$				$\hat{s}_3$			
	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR
t + MCMC	3.2	13.8	3.9	20.3	8.1	15.1	9.2	26.9	16.5	25.7	18.9	21.8
FMoG + EM	2.4	16.2	2.7	34.5	5.9	20.7	6.0	40.7	6.3	44.7	6.3	39.5

TABLE IV
ESTIMATES OF $\mathbf{A}$ FOR THE DETERMINED MIXTURE

Original matrix
$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 0.8 & 1.3 & -0.9 \\ 1.2 & -0.7 & 1.1 \end{bmatrix}$
t + MCMC
$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ 0.7938 & 1.3038 & -0.8978 \\ \pm(0.0046) & (\pm 0.0050) & (\pm 0.0049) \\ 1.1954 & -0.7039 & 1.0996 \\ \pm(0.0058) & \pm(0.0046) & \pm(0.0042) \end{bmatrix}$
FMoG + EM
$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ 0.7790 & 1.3181 & -0.8872 \\ 1.1864 & -0.7182 & 1.1069 \end{bmatrix}$
JADE
$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ 0.7814 & 1.3226 & -0.8974 \\ 1.3449 & -0.7920 & 0.8710 \end{bmatrix}$

$\mathbf{A}$ becomes invertible and $\boldsymbol{\mu}_{\tilde{\mathbf{s}}_k}$ is simply the result of the application of the (weighted) pseudoinverse of $\mathbf{A}$ to $\tilde{\mathbf{x}}_k$. When considering an overdetermined mixture with $m = 4$ sensors (with $\mathbf{r}_4^T = [-0.6\ 1.5\ 0.9]$ added to the rows of $\mathbf{A}$ given in Table IV), the SIRs of the source estimates rise, respectively, to 48.9, 50.0, and 46.2 dB with the $t + \mathrm{MCMC}$ method and to 44.1, 44.7, and 39.1 dB with the FMoG + EM method. However, the SARs obtained by both methods do not increase, which show the limitations of the priors studied.

Note the benefit of using sparsity for denoising purpose: the SNRs obtained by the methods $t + \mathrm{MCMC}$ and FMoG + EM are much higher than those obtained with JADE (JADE does not produce artifacts because the estimated sources are only a linear combination of the observations).

The mixing matrices are well estimated by the $t + \mathrm{MCMC}$ and FMoG + EM methods in both underdetermined and determined cases, with slightly better performance of the $t+\mathrm{MCMC}$ method. One advantage of the MCMC approach over EM is that standard variations of the estimated parameters can be computed easily.

2) Separation Performance With Respect to Mixing Matrix: Results of Section V-C emphasize another advantage of the MCMC approach compared with EM: robustness to mixing matrix. The $t + \mathrm{MCMC}$ method converges to the correct angle values in all four cases. On the contrary, the FMoG + EM approach only converges to the correct angles in case 1 and case 4. Several runs of both algorithms (with various random generator seeds for the MCMC method and several random initializations for the EM method) led to similar results. The problem encountered with the latter method is as follows. The clusters generated by the two close column directions in case 2 and case 3 are regarded as a single cluster: $\hat{\psi}_3$ is fitted between ψ_2 and ψ_3 in case 2, $\hat{\psi}_1$ is fitted between ψ_1 and ψ_2 in case 3. In these cases, the isolated direction is well estimated (ψ_1 in case 2 and ψ_3 in case 3) and the remaining direction of the estimated matrix is completely wrong.

However, the FMoG + EM method performs well in case 4 when all the columns of $\mathbf{A}$ are closely spread. It seems to be meaning that the problem encountered in case 2 and 3 is not only a problem of convergence of the algorithm to local modes, but also a matter of how the method "interprets" the data. As already mentioned in Section III, the better robustness of the MCMC approach derives from the fact that the algorithm explores all the space of possible values of $p(\boldsymbol{\theta} \mid \tilde{\mathbf{x}})$ whereas the EM approach only updates point estimates.

Interestingly, Table VII show that separation quality in case 1 is only slightly better than in case 4. In cases 2 and 3, the source along the isolated direction is very well estimated, much better than the 2 other sources, which SDRs surprisingly do not meet the values obtained in case 4.

3) Computational Issues and Convergence Properties: The MCMC approach has significantly higher computational burden than the EM approach, in particular the one designed in [11] which falls in the category of Alternating Expectation Conditional Maximization methods and, thus, shows faster convergence than the standard EM algorithm. In MATLAB, 1000 iterations of the $t+\mathrm{MCMC}$ method applied to a mixtures of $n = 3$ sources with $N = 65\,536$ require approximatively 3 h on a Mac G4 cadenced at 1.25 GHz, while it takes only a few minutes for the FMoG + EM method to converge. However, the FMoG + EM method is very sensitive to mixing matrix initialization and in our simulations the method happened to often converge to local minima. On the opposite, the $t + \mathrm{MCMC}$ method proved to be robust to initialization; the initialization to zeros of the rows of $\mathbf{A}$ to be estimated provided correct convergence, and other random initializations of $\mathbf{A}$ led to the same conclusions. However, the convergence toward the stationary distribution can be very slow and it might be safe to run multiple chains with different initialization of $\mathbf{A}$ [30, Ch. 6].

The faster convergence of the $t + \mathrm{MCMC}$ method for the estimation of the mixing parameters in the simulations concerning the determined case [Section V-B and Fig. 5(a) and (b)] compared to the simulations concerning the undetermined case [Section V-A and Fig. 3(a)] may be partially explained by the additional information given by the extra measurements.

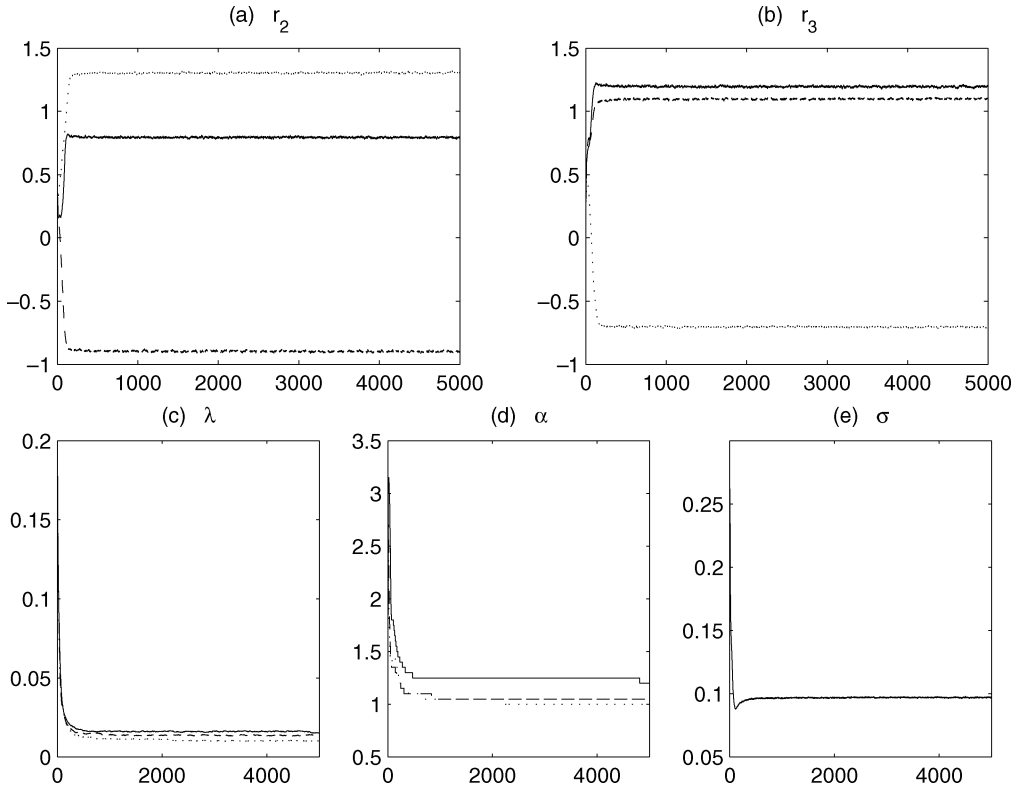

Fig. 5. Samples from the different parameters obtained with the Gibbs sampler for the determined mixture.

TABLE V
PERFORMANCE CRITERIA FOR THE DETERMINED MIXTURE

	$\hat{s}_1$				$\hat{s}_2$				$\hat{s}_3$			
	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR
t + MCMC	12.3	35.7	13.9	17.5	15.4	36.8	17.0	20.6	13.4	35.0	15.2	18.3
FMoG + EM	5.6	36.7	5.6	33.5	6.5	48.4	6.5	35.0	5.9	37.0	5.9	31.0
JADE	8.1	18.8	-	8.5	5.3	15.6	-	5.8	5.7	37.3	-	5.7

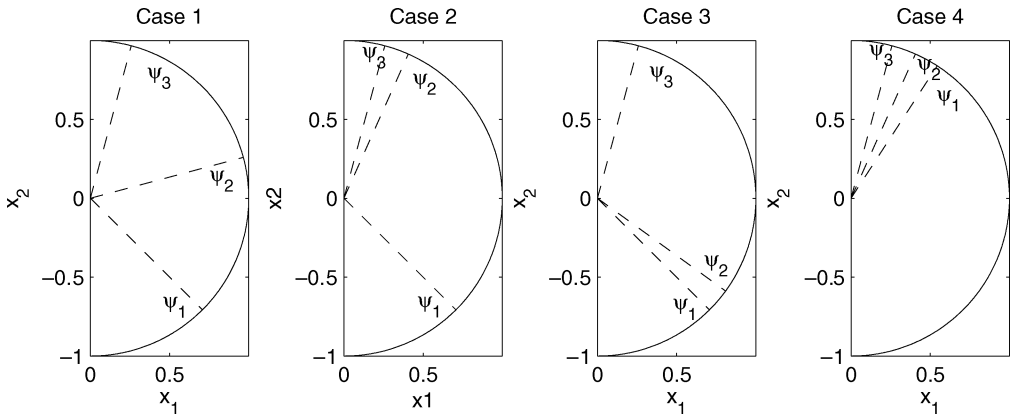

Fig. 6. Directions generated by the several matrices in the four cases studied.

However, another difference was the input noise variance σ. The higher the level of input noise, the less peaky the likelihood becomes; thus, the conditional posterior distributions of the parameters have a wider support and the Gibbs sampler "mixes" better [30]. When applying the Gibbs sampler to the underdetermined mixture presented in Section V-A with a lower noise level ($\sigma = 0.01$ instead of $\sigma = 0.03$), the estimate of $\mathbf{r}_2$ is slightly more accurate (with standard deviations divided by two) and the SNRs of the sources estimates increase by 7 dB, but convergence is only obtained after approximatively 4000 iterations (instead of approximatively 1500 iterations when $\sigma = 0.03$ as shown in Fig. 3).

Over 20 runs of the Gibbs sampler (corresponding to 20 different random generator seeds) on the underdetermined mixture

TABLE VI
ESTIMATED ANGLES

Case 1			
	ψ_1	ψ_2	ψ_3
True value (deg)	-45	15	75
t + MCMC	-44.6	16.4	75.3
FMoG + EM	-46.3	17.5	75.6
Case 2			
	ψ_1	ψ_2	ψ_3
True value (deg)	-45	66	75
t + MCMC	-43.9	66.2	75.4
FMoG + EM	-49.3	86.5	71.9
Case 3			
	ψ_1	ψ_2	ψ_3
True value (deg)	-45	-36	75
t + MCMC	-45.8	-35.6	75.2
FMoG + EM	-40.6	-83.6	75.2
Case 4			
	ψ_1	ψ_2	ψ_3
True value (deg)	57	66	75
t + MCMC	56.9	66.2	75.0
FMoG + EM	57.6	66.2	74.7

corresponding to Case 1 in Section V-C, 50% failed to converge to the stationary distribution within 5000 iterations when $\sigma = 0.03$, whereas all runs converged to the stationary distribution within 3000 iterations when $\sigma = 0.1$.

This might indicate that some kind of annealing strategy could improve convergence properties. Various possibilities exist for combining annealing with standard MCMC [29], parallel chain MCMC ([30, Ch. 6 and 7]) or annealed importance sampling [41]. In our framework, simulated annealing basically consists in artificially increasing the values taken by σ through the first iterations of the sampler and gradually decreasing them to their correct expected value. We propose a simple and effective way for annealing the target distribution. Preliminary results are presented with annealing of the degrees of freedom α_σ of the input noise: α_σ is replaced in (23) by $\alpha'_\sigma(k) = (1 - (1 - p_0)\exp(-k/k_0))\,\alpha_\sigma$, where k denotes the iteration. In this version of annealing, the degrees of freedom parameter for the input noise is exponentially increased to its correct value from a small starting value. In this way the sampler is more able to explore the probability distribution at earlier iterations, while effectively returning to the true stationary target distribution once $k \gg k_0$.

Fig. 7 shows a particular run of the Gibbs sampler applied to the data of Case 1 in Section V-C (where $\mathbf{r}_2^T = [-1\ 0.27\ 3.7]$ and $\sigma = 0.03$). In this particular case, the convergence is very slow, only reached after 10 000 iterations. Fig. 8 shows the same run of the Gibbs sampler (that is to say with the same random generator seeds) but with annealing of α_σ with $p_0 = 0.01$ and $k_0 = 300$. The acceleration of convergence is dramatic: a good estimation of the parameters can be obtained within hundreds of iterations instead of thousands without simulated annealing. The quality of the sources MMSE estimates (computed over the final 1000 samples in both cases) is the same with and without simulated annealing, with SDRs, SIRs, SARs, and SNRs similar to the figures for Case 1 in Table VII. There are clearly other ways to perform annealing in a setup such as this, and we have presented here only one possibility that works very well. In future work, we will provide further recommendations regarding optimal annealing setups for these problems.

We point out that another way to accelerate convergence is to initialize the mixing matrix with a rough estimate of $\mathbf{A}$, obtained with a clustering method like in [10], [35] or simply using the FMoG + EM method (provided it converges to the correct mixing matrix).

VI. CONCLUSION

We have presented a Bayesian approach to perform separation of linear instantaneous mixtures possibly underdetermined and noisy of sources having a sparse representation on a basis. We used a Student t distribution to model the sparsity of the coefficients. This distribution has the advantage that it can be expressed as a SMoG with an inverted-Gamma mixing weight function. This property allowed us to calculate easily the conditional posterior distributions required to design a Gibbs sampler. MMSE estimates of the mixing matrix and the sources were computed, and compared with an EM approach using a FMoG model with two states.

The primary aim of this paper is to show the good properties of the Student t prior for modeling sparsity. The MCMC approach is very robust but also computationally demanding. Possible ways to alleviate this would be EM approaches based on the Student t prior, in order to keep the increased audio quality with respect to the FMoG prior, while alleviating the computational burden implied the MCMC approach. However, this is likely to be at the expense of robustness to initialization and mixing matrix. We showed that simulated annealing schemes can dramatically improve the convergence of the sampler, optimized schemes are to be investigated. Convergence of the sampler may also be accelerated by initializing the mixing matrix to a rough estimate, which, when combined with simulated annealing and taking into account the robustness of the sampler to mixing matrix initialization and to mixing matrices close to singularity, is likely to challenge EM based methods.

Other perspectives are to extend our method to the case of overcomplete dictionaries, which would improve sparsity of the coefficients of the representation of the sources in the dictionary, and is likely to improve separation. Overcomplete lapped cosine transforms or Gabor dictionaries (as used in [15]) could be considered, as well as hybrid dictionaries such as unions of wavelets and MDCT bases (as used in [42]). Following the approach in [15], we can also consider using an indicator variable framework (i.e., a prior model of the form $p(\tilde{s}_{i,k} \mid \alpha_i, \lambda_i) = \gamma_{i,k}\,\delta_0(\tilde{s}_{i,k}) + (1 - \gamma_{i,k})\,t(\tilde{s}_{i,k} \mid \alpha_i, \lambda_i)$, with $\gamma_{i,k} \in \{0, 1\}$ and where δ_0 is Dirac's delta function). Besides further enforcing sparsity of the coefficients, this framework would easily allow to model audio signals time-frequency

TABLE VII
PERFORMANCE CRITERIA FOR THE DIFFERENT CASES

Case 1	$\hat{s}_1$				$\hat{s}_2$				$\hat{s}_3$			
	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR
t + MCMC	9.3	19.1	10.3	20.5	5.9	14.5	6.8	23.5	14.3	22.5	15.3	26.5
FMoG + EM	5.2	22.3	5.4	38.5	4.41	16.3	4.8	32.6	5.9	28.0	6.0	42.0

Case 2	$\hat{s}_1$				$\hat{s}_2$				$\hat{s}_3$			
	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR
t + MCMC	13.8	21.7	16.1	20.1	3.6	10.9	4.9	27.3	3.3	11.5	4.4	28.6
FMoG + EM	5.4	30.9	5.4	36.2	-30.9	-22.7	-7.5	40.7	-0.4	3.5	3.5	47.1

Case 3	$\hat{s}_1$				$\hat{s}_2$				$\hat{s}_3$			
	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR
t + MCMC	5.2	14.6	5.9	24.1	4.1	9.7	6.1	21.9	22.1	33.6	25.3	25.5
FMoG + EM	-0.4	0.3	10.2	36.0	-27.0	-15.7	-10.8	29.1	4.6	32.3	4.6	40.0

Case 4	$\hat{s}_1$				$\hat{s}_2$				$\hat{s}_3$			
	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR	SDR	SIR	SAR	SNR
t + MCMC	8.8	20.0	9.4	21.6	5.7	13.7	6.6	31.0	11.1	23.6	11.6	24.7
FMoG + EM	4.4	23.1	4.5	32.5	4.7	15.8	5.2	36.2	7.5	34.6	7.5	30.5

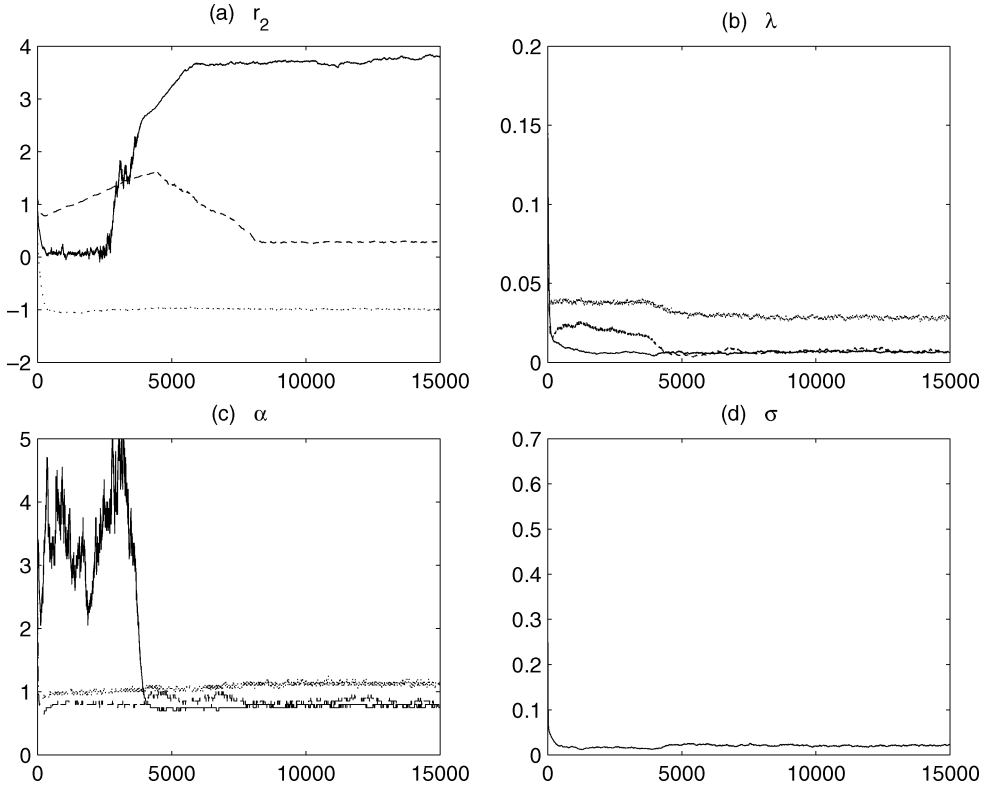

Fig. 7. One run of the Gibbs sampler on data corresponding to Case 1, without simulated annealing.

structures by choosing proper priors on γ to model harmonic structures.

Another useful direction will be making comparisons between the use of Student t prior and exponential power distributions family (to which the Laplacian belongs), which can also be expressed as SMoG [43], though the weight function does not lead to straightforward conditional distributions, but sampling could be achieved by rejection sampling schemes [30].

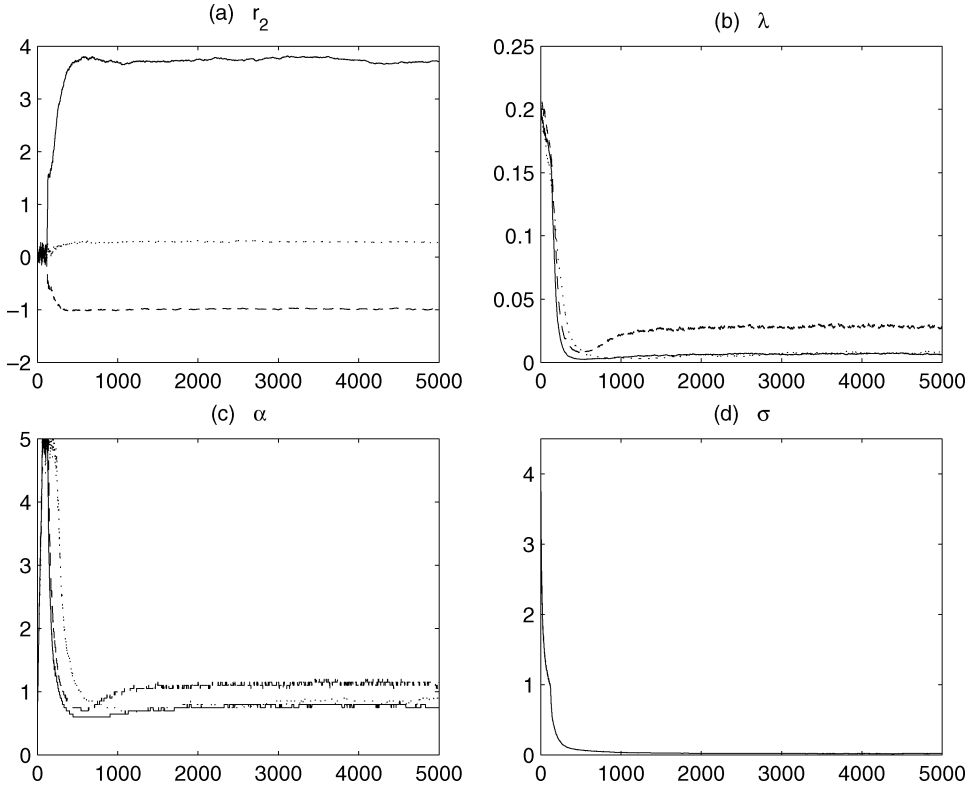

Fig. 8. One run of the Gibbs sampler on data corresponding to Case 1, with simulated annealing on α_σ: α_σ is replaced in (23) by $\alpha'_\sigma(k) = (1-(1-p_0)\,\exp(-k/k_0))\,\alpha_\sigma$, with in this case $p_0 = 0.01$ and $k_0 = 300$.

APPENDIX I
STANDARD DISTRIBUTIONS

Multivariate Gaussian: $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp -(1/2)(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. Gamma: $\mathcal{G}(x \mid \gamma,\beta) = (x^{\gamma-1}/\Gamma(\gamma)\,\beta^\gamma)\,\exp(-x/\beta)$. Inverted-Gamma: $\mathcal{IG}(x \mid \gamma,\beta) = (x^{-(\gamma+1)}/\Gamma(\gamma)\,\beta^\gamma)\,\exp(-1/\beta\,x)\mathbb{I}_{[0,+\infty)}(x)$. $\mathbb{I}_{[0,+\infty)}(x)$ denotes the indicator function of $[0,+\infty)$. The inverted-Gamma distribution is the distribution of $1/X$ when X is Gamma distributed.

APPENDIX II
DERIVATIONS OF CONDITIONAL DISTRIBUTIONS

A. Expression of $p(\tilde{\mathbf{s}} \mid \boldsymbol{\theta}_{-\tilde{\mathbf{s}}}, \tilde{\mathbf{x}})$

We have

$$p(\tilde{\mathbf{s}} \mid \boldsymbol{\theta}_{-\tilde{\mathbf{s}}}, \tilde{\mathbf{x}}) \propto p(\tilde{\mathbf{x}} \mid \mathbf{A}, \tilde{\mathbf{s}}, \sigma)\, p(\tilde{\mathbf{s}} \mid \mathbf{v}). \tag{30}$$

With $p(\tilde{\mathbf{s}} \mid \mathbf{v}) = \prod_{k=1}^N p(\tilde{\mathbf{s}}_k \mid \mathbf{v}_k) = \prod_{k=1}^N \mathcal{N}(\tilde{\mathbf{s}}_k \mid 0, \mathrm{diag}\,(\mathbf{v}_k))$ and with (20), we have

$$p(\tilde{\mathbf{s}} \mid \boldsymbol{\theta}_{-\tilde{\mathbf{s}}}, \tilde{\mathbf{x}}) \propto \prod_{k=1}^N \mathcal{N}(\tilde{\mathbf{x}}_k \mid \mathbf{A}\tilde{\mathbf{s}}_k, \sigma^2\mathbf{I}_m)\mathcal{N}(\tilde{\mathbf{s}}_k \mid 0, \mathrm{diag}\,(\mathbf{v}_k)). \tag{31}$$

The product of Gaussians for each k can be easily rearranged as another Gaussian function of $\tilde{\mathbf{s}}_k$, leading to (22).

B. Expression of $p(\mathbf{A} \mid \boldsymbol{\theta}_{-\mathbf{A}}, \tilde{\mathbf{x}})$

Let $\tilde{\mathbf{S}}_k$ and $\mathbf{a}$ denote the $m \times nm$ matrix and the $nm \times 1$ vector defined by

$$\tilde{\mathbf{S}}_k = \begin{bmatrix} \tilde{\mathbf{s}}_k^T & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \tilde{\mathbf{s}}_k^T \end{bmatrix} \quad \text{and} \quad \mathbf{a} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_m \end{bmatrix} \tag{32}$$

where we recall that $\mathbf{r}_1, \ldots, \mathbf{r}_m$ are the $n \times 1$ vectors denoting the transposed rows of $\mathbf{A}$. By construction, we have

$$\mathbf{A}\,\tilde{\mathbf{s}}_k = \tilde{\mathbf{S}}_k\,\mathbf{a}. \tag{33}$$

Of course, the estimation of $\mathbf{a}$ is equivalent to the estimation of $\mathbf{A}$, and we have

$$p(\mathbf{a} \mid \boldsymbol{\theta}_{-\mathbf{a}}, \tilde{\mathbf{x}}) \propto p(\tilde{\mathbf{x}} \mid \mathbf{a}, \tilde{\mathbf{s}}, \sigma)\, p(\mathbf{a}). \tag{34}$$

Without further information on the mixing matrix, we assume uninformative uniform prior on $\mathbf{A}$ and set $p(\mathbf{a}) \propto 1$. With (20) and (33), we then have $p(\mathbf{a} \mid \boldsymbol{\theta}_{-\mathbf{a}}, \tilde{\mathbf{x}}) \propto \prod_{k=1}^N \mathcal{N}(\tilde{\mathbf{x}}_k \mid \tilde{\mathbf{S}}_k\,\mathbf{a}, \sigma^2\,\mathbf{I}_m)$, which can be rearranged as the exponential of a quadratic form in $\mathbf{a}$, such that

$$p(\mathbf{a} \mid \boldsymbol{\theta}_{-\mathbf{a}}, \tilde{\mathbf{x}}) = \mathcal{N}(\mathbf{a} \mid \boldsymbol{\mu}_\mathbf{a}, \boldsymbol{\Sigma}_\mathbf{a}) \tag{35}$$

with $\boldsymbol{\Sigma}_\mathbf{a} = \sigma^2 \left(\sum_{k=1}^N \tilde{\mathbf{S}}_k^T \tilde{\mathbf{S}}_k\right)^{-1}$ and $\boldsymbol{\mu}_\mathbf{a} = (1/\sigma^2)\,\boldsymbol{\Sigma}_\mathbf{a} \sum_{k=1}^N \tilde{\mathbf{S}}_k^T\,\tilde{\mathbf{x}}_k$. It appears that $\boldsymbol{\Sigma}_\mathbf{a}$ is block-diagonal, and the rows of $\mathbf{A}$ are, thus, independent, leading, after simplifications, to (25).

C. Expression of $p(\sigma \mid \boldsymbol{\theta}_{-(\mathbf{A},\sigma)}, \tilde{\mathbf{x}})$

We have

$$p(\sigma \mid \boldsymbol{\theta}_{-(\mathbf{A},\sigma)}, \tilde{\mathbf{x}}) \propto p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta}_{-\mathbf{A}})\, p(\sigma) \tag{36}$$

and

$$p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta}_{-\mathbf{A}}) = \int_{\mathbf{A}} p(\tilde{\mathbf{x}}, \mathbf{A} \mid \boldsymbol{\theta}_{-\mathbf{A}}) d\mathbf{A} \tag{37}$$

$$= \int_{\mathbf{A}} p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta})\, p(\mathbf{A}) d\mathbf{A}. \tag{38}$$

Assuming uniform prior $p(\mathbf{A}) \propto 1$, the evaluation of $p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta}_{-\mathbf{A}})$, thus, requires integrating the likelihood over $\mathbf{A}$, or, in vector form, over $\mathbf{a}$. The likelihood can be written

$$p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{Nm/2}} \times \exp\left[-\frac{1}{2}\left((\mathbf{a}-\boldsymbol{\mu}_{\mathbf{a}})^T \boldsymbol{\Sigma}_{\mathbf{a}}^{-1}(\mathbf{a}-\boldsymbol{\mu}_{\mathbf{a}}) + c - \boldsymbol{\mu}_{\mathbf{a}}^T \boldsymbol{\Sigma}_{\mathbf{a}}^{-1} \boldsymbol{\mu}_{\mathbf{a}}\right)\right] \tag{39}$$

with $c = (1/\sigma^2)\sum_{k=1}^{N} \tilde{\mathbf{x}}_k^T\, \tilde{\mathbf{x}}_k$ and with $\boldsymbol{\Sigma}_{\mathbf{a}}$ and $\boldsymbol{\mu}_{\mathbf{a}}$ defined before. The exponential quadratic form in $\mathbf{a}$ integrating to $|2\,\pi\,\boldsymbol{\Sigma}_{\mathbf{a}}|^{1/2}$, it follows:

$$p(\tilde{\mathbf{x}} \mid \boldsymbol{\theta}_{-\mathbf{A}}) = \frac{|2\,\pi\,\boldsymbol{\Sigma}_{\mathbf{a}}|^{1/2}}{(2\,\pi\,\sigma^2)^{N\,m/2}} \exp -\frac{1}{2}\left(c - \boldsymbol{\mu}_{\mathbf{a}}^T\, \boldsymbol{\Sigma}_{\mathbf{a}}^{-1}\, \boldsymbol{\mu}_{\mathbf{a}}\right). \tag{40}$$

Expanding $\boldsymbol{\mu}_{\mathbf{a}}^T\, \boldsymbol{\Sigma}_{\mathbf{a}}^{-1}\, \boldsymbol{\mu}_{\mathbf{a}}$, grouping the terms in σ we obtain, with (36)

$$p(\sigma \mid \boldsymbol{\theta}_{-(\mathbf{A},\sigma)}, \tilde{\mathbf{x}}) \propto \left(\frac{1}{\sigma}\right)^{(N-n)\,m} \exp\left(-\frac{1}{\beta_\sigma\, \sigma^2}\right) p(\sigma) \tag{41}$$

where β_σ is defined at (24). With $p(\sigma) = 1/\sigma$, it is possible to show that σ has a square root inverted-Gamma distribution, (which is the distribution of $X^{-1/2}$ when X is Gamma distributed), leading to (23).

Expression of $p(\mathbf{v} \mid \boldsymbol{\theta}_{-\mathbf{v}}, \tilde{\mathbf{x}})$:

We have

$$p(\mathbf{v} \mid \boldsymbol{\theta}_{-\mathbf{v}}, \tilde{\mathbf{x}}) \propto p(\tilde{\mathbf{s}} \mid \mathbf{v})\, p(\mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\lambda}). \tag{42}$$

Owing to sources coefficients i.i.d. and mutual independence assumptions, we have

$$p(\tilde{\mathbf{s}} \mid \mathbf{v})\, p(\mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \prod_{k=1}^{N} \prod_{i=1}^{n} p(\tilde{s}_{i,k} \mid v_{i,k})\, p(v_{i,k} \mid \alpha_i, \lambda_i). \tag{43}$$

The product $p(\tilde{s}_{i,k} \mid v_{i,k})\, p(v_{i,k} \mid \alpha_i, \lambda_i)$ can be expressed straightforwardly (up to a constant) as an inverted-Gamma density function with parameters γ_{v_i} and $\beta_{v_{i,k}}$ defined in Section IV-D

D. Expressions of $p(\boldsymbol{\alpha} \mid \boldsymbol{\theta}_{-\boldsymbol{\alpha}}, \tilde{\mathbf{x}})$ and $p(\boldsymbol{\lambda} \mid \boldsymbol{\theta}_{-\boldsymbol{\lambda}}, \tilde{\mathbf{x}})$

We have

$$p(\boldsymbol{\alpha} \mid \boldsymbol{\theta}_{-\boldsymbol{\alpha}}, \tilde{\mathbf{x}}) \propto p(\mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\lambda})\, p(\boldsymbol{\alpha}) \tag{44}$$

$$p(\boldsymbol{\lambda} \mid \boldsymbol{\theta}_{-\boldsymbol{\lambda}}, \tilde{\mathbf{x}}) \propto p(\mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\lambda})\, p(\boldsymbol{\lambda}). \tag{45}$$

With $p(\mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\lambda})\, p(\boldsymbol{\alpha}) = \prod_{i=1}^{n} \left(\prod_{k=1}^{N} p(v_{i,k} \mid \alpha_i, \lambda_i)\right) p(\alpha_i)$ we get directly (23). Similarly, with $p(\mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\lambda})\, p(\boldsymbol{\lambda}) = \prod_{i=1}^{n} \left(\prod_{k=1}^{N} p(v_{i,k} \mid \alpha_i, \lambda_i)\right) p(\lambda_i)$, we get

$$p(\boldsymbol{\lambda} \mid \boldsymbol{\theta}_{-\boldsymbol{\lambda}}, \tilde{\mathbf{x}}) \propto \prod_{i=1}^{n} \lambda_i^{\alpha_i\, N} \exp\left(-\frac{\alpha_i\, R_i}{2}\, \lambda_i^2\right) p(\lambda_i). \tag{46}$$

With $p(\lambda_i) = 1/\lambda_i$, it is possible to show that λ_i has a square root-Gamma density (which is the distribution of $X^{1/2}$ when X is Gamma distributed), leading to (28).

ACKNOWLEDGMENT

The authors would like to thank M. Davies for the useful discussions about properties of the FMoG + EM method, as well as A. Doucet for his useful comments about this work. They would also like to thank L. Daudet for providing them with the MDCT code, as well as the anonymous reviewers for their valuable and encouraging comments.

REFERENCES

[1] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 9, no. 10, pp. 2009–2025, Oct. 1998.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[3] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique based on second order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[4] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[5] C. Févotte and C. Doncarli, "Two contributions to blind source separation using time-frequency distributions," *IEEE Signal Process. Lett.*, vol. 11, no. 3, pp. 386–389, Mar. 2004.

[6] B. A. Olshausen and K. J. Millman, "Learning sparse codes with a mixture-of-Gaussians prior," in *Advances in Neural Information Processing Systems*, S. A. Solla and T. K. Leen, Eds. Cambridge, MA: MIT Press, 2000, pp. 841–847.

[7] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, pp. 337–365, 2000.

[8] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Comput.*, vol. 13, no. 11, pp. 2517–2532, 2001.

[9] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Process. Lett.*, vol. 4, no. 4, pp. 87–90, Apr. 1999.

[10] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev, "Blind source separation by sparse decomposition," in *Independent Component Analysis: Principles and Practice*, S. J. Roberts and R. M. Everson, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[11] M. Davies and N. Mitianoudis, "A simple mixture model for sparse overcomplete ICA," presented at the IEE Conf. Vision, Image, and Signal Processing, Feb. 2004, unpublished.

[12] R. Gribonval, "Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture," presented at the ICASSP, Orlando, FL, May 2002, unpublished.

[13] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc. ICASSP*, Istanbul, Turkey, Jun. 2000, vol. 5, pp. 2985–2988.

[14] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.

[15] P. J. Wolfe, S. J. Godsill, and W.-J. Ng, "Bayesian variable selection and regularization for time-frequency surface estimation," *J. Roy. Stat. Soc. Ser. B*, vol. 66, pp. 575–590, 2004.

[16] D. Wipf, J. Palmer, and B. Rao, "Perspectives on sparse Bayesian learning," in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004, vol. 16.

[17] C. Févotte, S. J. Godsill, and P. J. Wolfe, "Bayesian approach for blind separation of underdetermined mixtures of sparse sources," in *Proc. 5th Int. Conf. Independent Component Analysis and Blind Source Separation*, Granada, Spain, 2004, pp. 398–405.
[18] K. H. Knuth, "Bayesian source separation and localization," in *Proc. SPIE Bayesian Inference for Inverse Problems*, San Diego, CA, Jul. 1998, pp. 147–158.
[19] ——, "A Bayesian approach to source separation," in *Proc. 1st Int. Workshop Independent Component Analysis and Signal Separation*, Aussois, France, Jan. 1999, pp. 283–288.
[20] K. H. Knuth and H. G. Vaughan, "Convergent Bayesian formulations of blind source separation and electromagnetic source estimation," in *Proc. Maximum Entropy and Bayesian Methods*, Munich, Germany, 1998, pp. 217–226.
[21] A. Mohammad-Djafari, "A Bayesian estimation method for detection, localization and estimation of superposed sources in remote sensing," presented at the SPIE, San Diego, CA, Jul. 1997, unpublished.
[22] ——, "A Bayesian approach to source separation," presented at the 19th Int. Workshop Bayesian Inference and Maximum Entropy Methods, Boise, ID, Aug. 1999, unpublished.
[23] D. B. Rowe, "A Bayesian approach to blind source separation," *J. Interdis. Math.*, vol. 5, no. 1, pp. 49–76, 2002.
[24] ——, *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*. Boca Raton, FL: CRC, 2003.
[25] S. Sénécal and P.-O. Amblard, "Bayesian separation of discrete sources via Gibbs sampling," in *Proc. International Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, 2000, pp. 566–572.
[26] C. Andrieu, A. Doucet, and S. J. Godsill, "Bayesian blind marginal separation of convolutively mixed discrete sources," presented at the IEEE Workshop Neural Networks for Signal Processing, Cambridge, MA, 1998, unpublished.
[27] C. Andrieu and S. J. Godsill, "Bayesian separation and recovery of convolutively mixed autoregressive sources," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 1999, pp. 1733–1736.
[28] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. Roy. Stat. Soc. Ser. B*, no. 36, pp. 99–102, 1974.
[29] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
[30] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall, 1996.
[31] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. Boston, MA: McGraw-Hill, 2002.
[32] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Boca Raton, FL: CRC, 2004, Texts in Statistical Science.
[33] J. S. Liu, "The collapsed Gibbs sampler with applications to a gene regulation problem," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 958–966, Sep. 1994.
[34] J. S. Liu, W. H. Wong, and A. Kong, "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes," *Biometrika*, vol. 81, no. 1, pp. 27–40, Mar. 1994.
[35] L. Vielva, I. Santamaria, D. Erdogmus, and J. C. Principe, "On the estimation of the mixing matrix for underdetermined blind source separation in an arbitrary number of dimensions," in *Proc. 5th Int. Conf. Independent Component Analysis and Blind Source Separation*, Granada, Spain, 2004, pp. 185–192.
[36] C. Liu and B. Rubin, "ML estimation of the t distribution using EM and its extensions, ECM and ECME," *Stat. Sin.*, vol. 5, pp. 19–39, 1995.
[37] K. Brandenburg, "MP3 and AAC explained," presented at the AES 17th Int. Conf. High Quality Audio Coding, Florence, Italy, Sep. 1999, unpublished.
[38] L. Daudet and M. Sandler, "MDCT analysis of sinusoids: Exact results and applications to coding artifacts reduction," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 302–312, May 2004.
[39] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," presented at the 4th Symp. Independent Component Analysis and Blind Source Separation, Nara, Japan, Apr. 2003, unpublished.
[40] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *Proc. Inst. Elect. Eng. F*, vol. 140, no. 6, pp. 362–370, 1993.
[41] R. M. Neal, "Annealed importance sampling," *Stat. Comput.*, vol. 11, pp. 125–139, 2001.
[42] S. Molla and B. Torrésani, "Determining local transientness in audio signals," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 625–628, Jul. .
[43] M. West, "On scale mixtures of normal distributions," *Biometrika*, vol. 74, no. 3, pp. 646–648, 1987.

Cédric Févotte was born in Laxou, France, in 1977, and lived in Tunisia, Senegal, and Madagascar until 1995. He received the Diplôme d'Études Appronfondies and the Diplôme de Docteur en Automatique et Informatique Appliquée from the l'École Centrale de Nantes and l'Université de Nantes, Nantes, France, in 2000 and 2003, respectively.

Since November 2003, he has been a Research Associate with the Signal Processing Laboratory, Engineering Department, University of Cambridge, Cambridge, U.K. His research interests concern statistical signal processing and time-frequency signal representations with application to blind source separation.

Simon J. Godsill (M'95) is Reader in statistical signal processing in the Engineering Department, University of Cambridge, Cambridge, U.K. He has research interests in Bayesian and statistical methods for signal processing, Monte Carlo algorithms for Bayesian problems, modeling and enhancement of audio and musical signals, and tracking and genomic signal processing. He has published extensively in journals, books, and conferences.

Dr. Godsill is an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING and the journal *Bayesian Analysis*, and is a member of IEEE Signal Processing Theory and Methods Committee. He co-edited a special issue of IEEE TRANSACTIONS ON SIGNAL PROCESSING on Monte Carlo methods in signal processing in 2002, and he has organized many conference sessions on related themes.