Bayesian audio source separation

Cédric Févotte

GET/Télécom Paris (ENST)
37-39, rue Dareau, 75014 Paris, France
E-mail: fevotte@tsi.enst.fr

Abstract

In this chapter we describe a Bayesian approach to audio source separation. The approach relies on probabilistic modeling of sound sources as (sparse) linear combinations of atoms from a dictionary and Markov chain Monte Carlo (MCMC) inference. Several prior distributions are considered for the source expansion coefficients. We first consider independent and identically distributed (iid) general priors with two choices of distributions. The first one is the Student t, which is a good model for sparsity when the shape parameter has a low value. The second one is a hierarchical mixture distribution; conditionally upon an indicator variable, one coefficient is either set to zero or given a normal distribution, whose variance is in turn given an inverted-Gamma distribution. Then, we consider more audio-specific models where both the identically distributed and independently distributed assumptions are lifted. Using a Modified Discrete Cosine Transform (MDCT) dictionary, a time-frequency orthonormal basis, we describe frequency-dependent structured priors which explicitly model the harmonic structure of sound, using a Markov hierarchical modeling of the expansion coefficients. Separation results are given for a stereophonic recording of 3 sources.

1 Introduction

In this chapter we take a Bayesian approach to blind source separation (BSS). We limit our study to the linear instantaneous problem, possibly underdetermined. Our notations are such that, for $t = 1, \ldots, N$

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

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 $\mathbf{e}_t = [e_{1,t}, \ldots, e_{m,t}]^T$ is a vector of size m containing additive noise/residual error. Variables without time index t will denote whole sequences of samples, *e.g*, $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and $x_1 = [x_{1,1}, \ldots, x_{1,N}]$. $\mathbf{A}$ is the mixing matrix of size $m \times n$ and all the latter variables are assumed to be real-valued.

1.1 The Bayesian setting

One fundamental principle of the Bayesian approach is to consider all the unknown variables of the problem $\mathbf{A}$, $\mathbf{s}$, $\mathbf{e}$ to be the realizations of random variables. This approach hence underlies a *probabilistic modeling* of these parameters, as opposed to have a fixed, ground-truth value. Each of these variable is assumed to follow a model described by a set of parameters respectively denoted by $\boldsymbol{\theta}_\mathbf{A}$, $\boldsymbol{\theta}_\mathbf{s}$ and $\boldsymbol{\theta}_\mathbf{e}$. The parameters $\mathbf{A}$, $\mathbf{s}$, $\mathbf{e}$ are thus characterized by their *prior distributions* $p(\mathbf{A}|\boldsymbol{\theta}_\mathbf{A})$, $p(\mathbf{s}|\boldsymbol{\theta}_\mathbf{s})$ and $p(\mathbf{e}|\boldsymbol{\theta}_\mathbf{e})$. The parameters $\boldsymbol{\theta}_\mathbf{A}$, $\boldsymbol{\theta}_\mathbf{s}$ and $\boldsymbol{\theta}_\mathbf{e}$ are referred to as *hyperparameters*. Depending on the degree of prior knowledge about $\mathbf{A}$, $\mathbf{s}$ and $\mathbf{e}$ the hyperparameters can be either fixed or treated as unknown parameters themselves, to be estimated (or, in a machine learning parlance, to be *learnt*) from the data $\mathbf{x}$. In the following, the set of all the parameters is noted $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{s}, \boldsymbol{\theta}_\mathbf{A}, \boldsymbol{\theta}_\mathbf{s}, \boldsymbol{\theta}_\mathbf{e}\}$.[1] Fig. 1 gives a graph representation of the model and shows the connections between the various parameters.

[1] $\mathbf{e}$ is omitted of the set as we simply have $\mathbf{e} = \mathbf{x} - \mathbf{A}\,\mathbf{s}$.

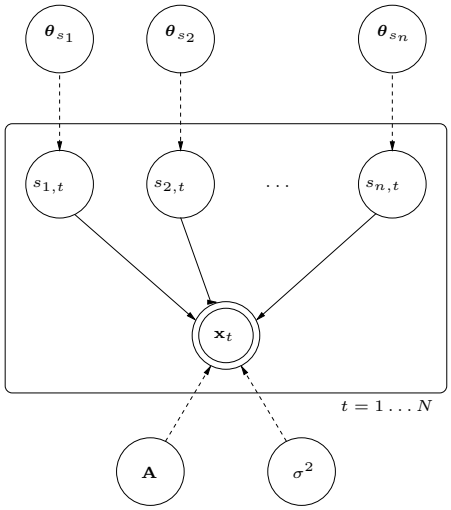

Figure 1: Directed acyclic graph (DAG) representation of the linear instantaneous BSS problem.

Given this statistical setting of the problem, Bayesian estimation revolves around the *posterior distribution* of the set of all unknown parameters $p(\boldsymbol{\theta}|\mathbf{x})$. Information about $\boldsymbol{\theta}$ or subsets of $\boldsymbol{\theta}$ is *inferred* from the data through manipulation of the posterior. As such, typical *point estimates* are the maximum a posteriori (MAP) estimate $\hat{\boldsymbol{\theta}}_{MAP} = \text{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\mathbf{x})$ and the minimum mean square error (MMSE) estimate $\hat{\boldsymbol{\theta}}_{MMSE} = \text{E}\{\boldsymbol{\theta}|\mathbf{x}\} = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}$.

A wide literature now exists about Bayesian source separation, covering many source and mixture models, as well as many computational techniques. Early works include the ones of Knuth [1, 2] and Djafari [3]. These authors show how some standard independent component analysis (ICA) techniques (such as InfoMax [4]) can easily be rephrased in a Bayesian parlance, but also how the Bayesian formalism easily allows to go beyond the standard hypotheses of ICA - square mixing matrix, independent and identically distributed (iid) sources, no additive noise/residual error. As such, Knuth shows in [1, 2] how the Bayesian formalism allows to incorporate prior knowledge about the location of the sensors (the mixing matrix) in the separation process. In [3], Djafari presents MAP gradient approaches for separation of possibly underdetermined and noisy mixtures of possibly spatially correlated and non-iid sources.

Bayesian approaches are precisely of utter interest in the underdetermined case. In the overdetermined non-noisy case, it is sufficient to estimate the mixing matrix and to apply its inverse to the observations to reconstruct the sources. In that case Cardoso shows in [5] that the sources need not to be accurately modeled to obtain a good estimation of the mixing matrix, and thus a good separation. Oppositely, the underdetermined case is an ill-posed problem because the mixing matrix is not invertible, and prior information about the sources is important to their reconstruction (the more information is available, the better the reconstruction). Prior information is also required in noisy scenarios. The Bayesian formalism is adapted to these problems because it allows to gather the available information about the sources in the prior $p(\mathbf{s}|\boldsymbol{\theta}_\mathbf{s})$. Note that there is a trade-off between the complexity of the source models and the complexity of the inference step to follow (consisting of minimizing a criterion or sampling a distribution). The source models should be built as to gather as much information as possible while keeping the inference step tractable.

1.2 Sparse models: analysis vs synthesis

In this chapter, we address Bayesian separation of underdetermined noisy mixtures of audio sources. The source models that we use are based on *sparsity* properties. A sequence is said to be sparse when most of its samples take a value close to zero. The use of sparsity to handle the general linear instantaneous model, has arisen in several papers in the areas of learning [6, 7, 8] and source separation [9, 10, 11, 12] (to name a few). For source separation, a linear transform is typically

applied to each observation, transposing the problem in a *transform domain* where the sources become sparse. Most of the latter papers then take a Bayesian approach: the sources are assumed to be the realizations of a iid random process with distribution gathering most of its probability around zero and presenting heavy tails, thus modeling sparsity. This distribution is the Laplace distribution in [7, 8, 9, 10], a mixture of Gaussians in [6], a mixture of Gaussian and Dirac in [11], a generalized Gaussian in [13]. Various inference strategies follow, all gradient-based or using the Expectation Maximization (EM) algorithm, and aiming at MAP or MMSE estimates of $\mathbf{A}$ and the sources.

The common ingredient of these methods is to work in a transform domain. There are two dual approaches for doing so. In many of the latter papers, a short-time Fourier transform is typically *applied* to the observations when dealing with audio, while a wavelet transform is used for images. This is an *analysis* approach. Given a dictionary $\mathbf{\Phi} \in \mathbb{R}^{K \times N}$ containing K waveforms, also referred to as atoms (with $K \geq N$), the analysis approach consists of computing the dot products of every atom of the dictionary with each observation, such that, $\forall j = 1, \ldots, m$

$$\tilde{x}_j^{an} = x_j \, \mathbf{\Phi}^T \tag{2}$$

and yielding

$$\tilde{\mathbf{x}}^{an} = \mathbf{A} \, \tilde{\mathbf{s}}^{an} + \tilde{\mathbf{e}}^{an} \tag{3}$$

where $\tilde{\mathbf{x}}^{an} = \mathbf{x} \, \mathbf{\Phi}^T$, $\tilde{\mathbf{s}}^{an} = \mathbf{s} \, \mathbf{\Phi}^T$ and $\tilde{\mathbf{e}}^{an} = \mathbf{e} \, \mathbf{\Phi}^T$. The analysis approach thus simply transforms the time domain linear instantaneous mixture in another linear instantaneous mixture. Any BSS method can be applied to the new mixture, and when an estimate $\widehat{\tilde{\mathbf{s}}^{an}}$ is obtained, a time domain estimate can be reconstructed through the pseudo-inverse of $\mathbf{\Phi}^T$ or, in a *frame* parlance, the dual operator [14], such that

$$\widehat{\mathbf{s}} = \widehat{\tilde{\mathbf{s}}^{an}} \; \mathbf{\Phi} \, (\mathbf{\Phi}^T \, \mathbf{\Phi})^{-1} \tag{4}$$

Conversely, the *synthesis* approach models the sources as a linear combination of atoms from $\mathbf{\Phi}$, such that $\forall i = 1, \ldots, n$

$$s_i = \tilde{s}_i^{sy} \, \mathbf{\Phi} \tag{5}$$

and yielding

$$\mathbf{x} = \mathbf{A} \, \tilde{\mathbf{s}}^{sy} \, \mathbf{\Phi} + \mathbf{e} \tag{6}$$

The two approaches, analysis and synthesis, differ very much in nature. The analysis approach aims at *sparsifying* the data, and its underlying sources, while the synthesis approach intrinsically models the sources as a sparse linear combination of atoms. The motivation of the analysis approach is to come up with sources with lower entropy, which tends to cluster the data along the mixing matrix columns (see Chapter 7), and potentially yields estimates with lower error variance [5]. The synthesis approach is *generative* in essence, it allows to build source models taking into account the specificities of the physical phenomena which generated the signals. Note that, when $s = \tilde{s}^{sy} \, \mathbf{\Phi}$ and $\mathbf{\Phi}$ is overcomplete ($K > N$), the synthesis coefficients $\tilde{s}^{sy}$ are not retrieved through direct analysis, as we have $\tilde{s}^{an} = \tilde{s}^{sy} \, \mathbf{\Phi} \, \mathbf{\Phi}^T \neq \tilde{s}^{sy}$. Furthermore, the operator $\mathbf{\Phi} \, \mathbf{\Phi}^T$ creates, in general, a "blurring" effect and $\tilde{s}^{an}$ is potentially less sparse than $\tilde{s}^{sy}$ [15].

The *analysis* and *synthesis* coefficients coincide when $\mathbf{\Phi}$ is an orthonormal basis, i.e, $\mathbf{\Phi} \, \mathbf{\Phi}^T = \mathbf{I}_N$, which is the case we address in this chapter. We thus note $\tilde{s}_i = \tilde{s}_i^{an} = \tilde{s}_i^{sy}$. However, we will abide to a synthesis interpretation of the models described, and will mention the generalization to the overcomplete case when possible.

This chapter is organized as follows. In Section 2 we describe a general Bayesian framework for blind separation of sparse sources. Two source priors are considered in Section 2.1: the Student t prior and a hierarchical mixture prior. The first prior is a distribution with two parameters, one scale parameter and one shape parameter controlling the peakiness of the density (and thus the degree of sparsity). We will take advantage of its hierarchical formulation as a scale mixture of Gaussians (SMoG), which yields computational facilities. The second prior has a hierarchical structure too, it consists of a mixture of a Dirac distribution centered at 0 and a normal distribution, whose variance is in turn given a conjugate inverted-Gamma prior. We then derive in Section 2.3 a Gibbs sampler, a standard Markov chain Monte Carlo method, to generate samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$.

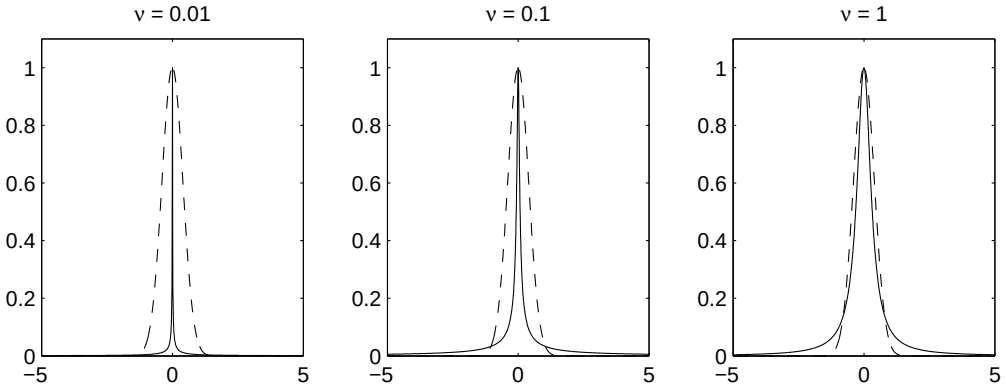

Figure 2: Student t densities for $\nu \in \{0.01, 0.1, 1\}$ with equal value at the mode. The dash-lined plot is the Gaussian density with variance $1/2\,\pi$.

While the methodology presented in Section 2 is very general, in Section 3 we more specifically address audio sources. We describe improvements to the latter priors in order to take into account the specificities of audio signals. As such, using a MDCT time-frequency orthonormal basis, we show how to model simply the non-uniform energy distribution of audio signals and how to model structural constraints (corresponding to physical phenomena such as tonal parts). We present in Section 4 results of the separation of a stereophonic musical recording with three sources (singing voice, acoustic guitar, bass guitar) using the techniques described. Finally, conclusions are given in Section 5.

2 Bayesian sparse source separation

2.1 Sparse source priors

In this section, we model each source coefficients sequence $\tilde{s}_i$ as an iid sequence, with two distributions considered, the Student t distribution and a hierarchical mixture prior. The sequences $\tilde{s}_1, \ldots, \tilde{s}_n$ are furthermore modeled as mutually independent, such that the prior distribution $p(\tilde{\mathbf{s}}|\boldsymbol{\theta}_{\tilde{\mathbf{s}}})$ factorizes as

$$p(\tilde{\mathbf{s}}|\boldsymbol{\theta}_{\mathbf{s}}) = \prod_{i=1}^{n} \prod_{k=1}^{N} p(\tilde{s}_{i,k}|\boldsymbol{\theta}_{s_i}) \tag{7}$$

2.1.1 Student t prior

The Student t prior is given by

$$p(\tilde{s}_{i,k}|\nu_i, \xi_i) = \frac{\Gamma(\frac{\nu_i+1}{2})}{\xi_i \sqrt{\nu_i\,\pi}\,\Gamma(\frac{\nu_i}{2})} \left(1 + \frac{1}{\nu_i}\left(\frac{\tilde{s}_{i,k}}{\xi_i}\right)^2\right)^{-\frac{\nu_i+1}{2}} \tag{8}$$

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

The Student t distribution can be expressed as a SMoG [16], such that

$$p(\tilde{s}_{i,k}|\nu_i, \xi_i) = \int_0^{+\infty} \mathcal{N}(\tilde{s}_{i,k}|0, v_{i,k})\,\mathcal{IG}\left(v_{i,k}|\alpha_i, \lambda_i\right) dv_{i,k} \tag{9}$$

where

$$\alpha_i = \frac{\nu_i}{2} \quad \lambda_i = \frac{\nu_i \, \xi_i^2}{2} \tag{10}$$

and where $\mathcal{N}(x|\mu, v)$ and $\mathcal{IG}(x|\alpha, \beta)$ are the Gaussian and inverted-Gamma distributions, defined in Appendix A.1. The variance $v_{i,k}$ can be treated as an auxiliary variable and $p(\tilde{s}_{i,k}|\nu_i, \xi_i)$ can be thus interpreted as a marginal density of the joint distribution $p(\tilde{s}_{i,k}, v_{i,k}|\nu_i, \xi_i)$, defined by

$$p(\tilde{s}_{i,k}, v_{i,k}|\nu_i, \xi_i) = p(\tilde{s}_{i,k}|v_{i,k}) \, p(v_{i,k}|\alpha_i, \lambda_i) \tag{11}$$

with

$$p(\tilde{s}_{i,k}|v_{i,k}) = \mathcal{N}(\tilde{s}_{i,k}|0, v_{i,k}) \quad \text{and} \quad p(v_{i,k}|\alpha_i, \lambda_i) = \mathcal{IG}\left(v_{i,k}|\alpha_i, \lambda_i\right) \tag{12}$$

This hierarchical formulation of the Student t, and the fact that the prior of $v_{i,k}$ is conjugate,[2] lead to easy Gibbs updates for both $\tilde{s}_{i,k}$ and $v_{i,k}$ as shown in Section 2.3. A graphical representation of the Student t source coefficient model is given at Fig. 3.

The source hyperparameters $\boldsymbol{\theta}_{s_i} = \{\alpha_i, \lambda_i\}$ can be given priors as well. Low values of ν_i, supporting sparsity, can be favored using and exponential prior

$$p(\alpha_i) = \beta_{\alpha_i} \, \exp(-\beta_{\alpha_i} \, \alpha_i). \tag{13}$$

A Gamma conjugate prior is chosen for λ_i, such that

$$p(\lambda_i) = \mathcal{G}(\lambda_i|\alpha_{\lambda_i}, \beta_{\lambda_i}) \tag{14}$$

The Student t can be interpreted as an infinite sum of Gaussians, which contrasts with the finite sums of Gaussians used in [6, 11]. The Laplace prior used in [9, 7] can also be expressed as a SMoG, with an exponential density on the variance $v_{i,k}$ [16]. However, the Student t prior has the advantage to offer a supplementary hyperparameter α_i which controls the sharpness of the distribution. The Laplace and the Student t belong to the more general family of general hyperbolic processes, with many shape parameters, which is used for source separation in [17]. However, the authors of [17] point out the limits of this class of priors, which is too general: too much flexibility does not in the end bring any prior information.

2.1.2 Hierarchical mixture prior

The second prior we consider is a hierarchical mixture prior, given by

$$\begin{aligned} p(\tilde{s}_{i,k}|\gamma_{i,k}, v_{i,k}) &= (1 - \gamma_{i,k}) \, \delta_0(\tilde{s}_{i,k}) + \gamma_{i,k} \, \mathcal{N}(\tilde{s}_{i,k}|0, v_{i,k}) & (15) \\ p(v_{i,k}|\alpha_i, \lambda_i) &= \mathcal{IG}(v_{i,k}|\alpha_i, \lambda_i) & (16) \end{aligned}$$

where $\delta_0(u)$ is the Dirac delta function and $\gamma_{i,k} \in \{0, 1\}$ is an indicator variable. When $\gamma_{i,k} = 0$, $\tilde{s}_{i,k}$ is set to zero; when $\gamma_{i,k} = 1$, $\tilde{s}_{i,k}$ has a normal distribution with zero mean and variance $v_{i,k}$, which is in turn assigned a conjugate inverted-Gamma prior. The set of indicator variables γ_i is so far modeled as iid, with Bernoulli prior

$$P(\gamma_{i,k} = 1|P_i) = P_i \quad P(\gamma_{i,k} = 0|P_i) = 1 - P_i \tag{17}$$

A graphical representation of the model is given at Fig. 3.

Contrary to the Student t, this prior *explicitly* models the possibility for a coefficient to be zero. Note that, conditionally upon $\gamma_{i,k} = 1$, the marginal $p(\tilde{s}_{i,k}|\gamma_k = 1, \alpha_i, \lambda_i)$ is Student t, with density $t(\tilde{s}_{i,k}|2\,\alpha_i, \sqrt{\lambda_i/\alpha_i})$. However, the degrees of freedom do not here play a key role like in previous paragraph, and in particular, it does not need to be given a small value. Here, the degree of sparsity is controlled by the parameter P_i, i.e the probability of one coefficient to be nonzero.

[2] If a parameter θ is observed through the observation x via the likelihood $p(x|\theta)$, the prior $p(\theta)$ is said to be conjugate when $p(\theta)$ and $p(\theta|x) \propto p(x|\theta)\, p(\theta)$ belong to the same family of distributions. Here, $v_{i,k}$ is observed through $\tilde{s}_{i,k}$ via $p(\tilde{s}_{i,k}|v_{i,k})$, its prior is $\mathcal{IG}(v_{i,k}|\alpha_i, \lambda_i)$ and its posterior $p(v_{i,k}|\tilde{s}_{i,k}, \alpha_i, \lambda_i)$, given at (38) is also inverted-Gamma.

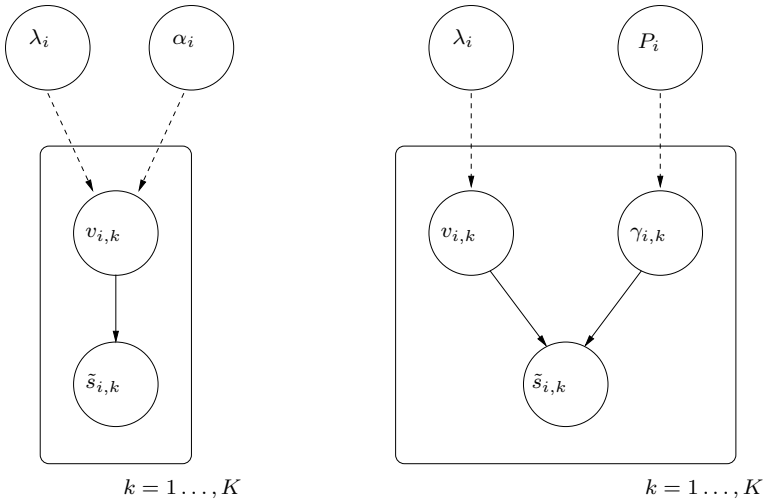

Figure 3: Graphs of the source coefficients models. Left: Student t; Right: hierarchical mixture prior.

The value of α_i is fixed to a value chosen by hand (see Section 4). Note that with the iid Bernoulli prior, $\gamma_{i,k}$ could be integrated out from (15), yielding simply

$$p(\tilde{s}_{i,k}|\boldsymbol{\theta}_{s_i}) = (1 - P_i)\,\delta_0(\tilde{s}_{i,k}) + P_i\,t(\tilde{s}_{i,k}|2\,\alpha_i, \sqrt{\lambda_i/\alpha_i}) \tag{18}$$

with $\theta_{s_i} = \{P_i, \lambda_i\}$. Like $v_{i,k}$, $\gamma_{i,k}$ is an auxiliary variable which, in the iid case, could be easily removed from the notations. It will play a more important role in Section 3.3 when used to model dependencies (structures) within the set of coefficients $\tilde{s}_i$.

The scale parameter λ_i is given a Gamma conjugate prior as before (see (14)). The probability P_i is given a Beta prior (defined in Appendix A.1), such that

$$p(P_i) = \mathcal{B}(P_i|\alpha_{P_i}, \beta_{P_i}) \tag{19}$$

The values of α_{P_i} and β_{P_i} can be adjusted as to yield a prior for P_i favoring low values and thus sparsity.

2.2 Noise and mixing matrix priors

We assume for simplicity the sequence $\mathbf{e}$ to be iid Gaussian, with covariance $\sigma^2\,\mathbf{I}_m$. However, different noise variances on every observation x_j could also easily be considered. When an orthonormal basis is used (*i.e*, $\boldsymbol{\Phi}^{-1} = \boldsymbol{\Phi}^T$), $\tilde{\mathbf{e}}$ is equivalently iid Gaussian with covariance $\sigma^2\,\mathbf{I}_m$, so that we have equivalence between time domain ($\mathbf{x}_t = \mathbf{A}\,\mathbf{s}_t + \mathbf{e}_t$) and transform domain ($\tilde{\mathbf{x}}_k = \mathbf{A}\,\tilde{\mathbf{s}}_k + \tilde{\mathbf{e}}_k$).

The variance σ^2 is treated as an unknown parameter ($\boldsymbol{\theta}_{\mathbf{e}} = \{\sigma^2\}$), and can be given a conjugate inverted-Gamma prior, such that

$$p(\sigma^2) = \mathcal{IG}(\sigma^2|\alpha_{\sigma^2}, \beta_{\sigma^2}) \tag{20}$$

The matrix $\mathbf{A}$ is treated in the following as a column vector $\mathbf{a}$ defined by

$$\mathbf{a} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_n \end{bmatrix} \tag{21}$$

where $\mathbf{r}_1, \ldots, \mathbf{r}_n$ denote the transposed rows of $\mathbf{A}$. A Gaussian conjugate prior could be used for $\mathbf{a}$, however, in practice we will simply use a noninformative flat prior $p(\mathbf{a}) \propto 1$, so that $\boldsymbol{\theta}_{\mathbf{A}}$ is empty.

Algorithm 1 Gibbs sampler

Initialize $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_M^{(0)}\}$
for $l = 1 : L + L_{bi}$ **do**
  $\boldsymbol{\theta}_1^{(l)} \sim p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(l-1)}, \ldots, \boldsymbol{\theta}_M^{(l-1)}, \mathbf{x})$
  $\boldsymbol{\theta}_2^{(l)} \sim p(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(l)}, \boldsymbol{\theta}_3^{(l-1)}, \ldots, \boldsymbol{\theta}_M^{(l-1)}, \mathbf{x})$
  $\boldsymbol{\theta}_3^{(l)} \sim p(\boldsymbol{\theta}_3 | \boldsymbol{\theta}_1^{(l)}, \boldsymbol{\theta}_2^{(l)}, \boldsymbol{\theta}_4^{(l-1)} \ldots, \boldsymbol{\theta}_M^{(l-1)}, \mathbf{x})$
  $\vdots$
  $\boldsymbol{\theta}_M^{(l)} \sim p(\boldsymbol{\theta}_M | \boldsymbol{\theta}_1^{(l)}, \boldsymbol{\theta}_2^{(l)}, \ldots, \boldsymbol{\theta}_{M-1}^{(l)}, \mathbf{x})$
end for

2.3 Markov chain Monte Carlo inference

We derive in the following a Gibbs sampler to generate samples (realizations) from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$. The obtained samples $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(L)}\}$ then allow for computation of any required point estimate (as well as interval estimates). For example, the Minimum Mean Square Error (MMSE) estimate is approximated by

$$\hat{\boldsymbol{\theta}}_{MMSE} \approx \frac{1}{L} \sum_{l=1}^{L} \boldsymbol{\theta}^{(l)} \tag{22}$$

The Gibbs sampler only requires to be able to sample from the posterior distribution of certain subsets of parameters conditional upon data $\mathbf{x}$ and the remaining parameters [18, 19]. Let $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ denote a partition of $\boldsymbol{\theta}$. The Gibbs sampler is described by Algorithm 1. L_{bi} represents the number of iterations required before the Markov chain $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots\}$ reaches its stationary distribution $p(\boldsymbol{\theta}|\mathbf{x})$ (the *burn in* period). Thereafter, all samples are drawn from the desired stationary distribution. MCMC methods have the advantage to generate samples from the whole support of $p(\boldsymbol{\theta}|\mathbf{x})$, 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 theoretically 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 MAP estimates) directly.

The implementation of a Gibbs sampler thus requires to define a partition of the set of parameters $\boldsymbol{\theta}$, and then to sample each subset conditionally upon the others and the data. We will discuss different ways to partition $\boldsymbol{\theta}$ in the following, but as a general rule, fastest convergence is obtained when as many parameters as possible are sampled jointly [20, 21]. In that sense, Gibbs sampling can be thought of as a stochastic version of iterated relaxed gradient optimization where deterministic moves are replaced by random moves. The largest moves in the right direction, the fastest convergence.

2.3.1 How to write a conditional distribution ?

We note $\boldsymbol{\theta}_{-j}$ the set $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, \ldots, \boldsymbol{\theta}_M\}$. We need to sample from $p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j}, \mathbf{x})$, $\forall j = 1, \ldots, M$. The conditional distribution of $\boldsymbol{\theta}_j$ is defined by

$$p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j}, \mathbf{x}) = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{p(\boldsymbol{\theta}_{-j}|\mathbf{x})} \tag{23}$$

Bayes' theorem gives

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{x})} \tag{24}$$

and it follows

$$p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j}, \mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta}_j, \boldsymbol{\theta}_{-j})\, p(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{-j})}{p(\boldsymbol{\theta}_{-j}|\mathbf{x})\, p(\mathbf{x})} \tag{25}$$

$$\propto \underbrace{p(\mathbf{x}|\boldsymbol{\theta}_j, \boldsymbol{\theta}_{-j})}_{likelihood} \underbrace{p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j})}_{prior} \tag{26}$$

In our case the likelihood is Gaussian and the prior $p(\boldsymbol{\theta})$ factorizes as

$$p(\boldsymbol{\theta}) = p(\mathbf{A})\, p(\sigma^2) \prod_{i=1}^{n} p(s_i|\boldsymbol{\theta}_{s_i}) \tag{27}$$

In the following we only discuss the ways to partition $\boldsymbol{\theta}$ and skip the details about how to calculate the conditional posterior distribution. Since most of the priors used are conjugate, the posteriors are rather straightforward to obtain, and belong to families of distributions easy to sample from, namely unidimensional or multivariate Gaussian and (inverted-)Gamma distributions. Further details can be found in [22].

2.3.2 Update of A and σ^2

Fig. 1 shows that the conditional posterior distributions of $\mathbf{A}$ and σ^2 merely depend on the source signals $\mathbf{s}$ (and independently of how they are modeled) and $\mathbf{x}$.

When the flat prior $p(\mathbf{a}) \propto 1$ is used, the rows of $\mathbf{A}$ are found to be a posteriori mutually independent with

$$p(\mathbf{r}_i|\sigma^2, \mathbf{s}, \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{r}_i}, \boldsymbol{\Sigma}_{\mathbf{r}}) \tag{28}$$

where $\boldsymbol{\Sigma}_{\mathbf{r}} = \sigma^2 \left(\sum_t \mathbf{s}_t\, \mathbf{s}_t^T\right)^{-1}$ and $\boldsymbol{\mu}_{\mathbf{r}_i} = \frac{1}{\sigma^2}\, \boldsymbol{\Sigma}_{\mathbf{r}} \sum_t x_{i,t}\, \mathbf{s}_t$.[3] The posterior distribution of σ^2 is simply

$$p(\sigma^2|\mathbf{A}, \mathbf{s}, \mathbf{x}) = \mathcal{IG}(\sigma^2|\alpha_{\sigma^2}^{post}, \lambda_{\sigma^2}^{post}) \tag{29}$$

with

$$\alpha_{\sigma^2}^{post} = \frac{N\, m}{2} + \alpha_{\sigma^2} \quad \text{and} \quad \beta_{\sigma^2}^{post} = \frac{1}{2} \|\mathbf{x} - \mathbf{A}\, \mathbf{s}\|_F^2 + \beta_{\sigma^2} \tag{30}$$

Note that $\alpha_{\sigma^2}^{post}$ and $\beta_{\sigma^2}^{post}$ reflects the trade-off between information provided from observation and from the prior. When N is large, α_{σ^2} and β_{σ^2} have little influence in the posterior. As such, Jeffreys noninformative prior $p(\sigma^2) \propto 1/\sigma^2$, corresponding to $\alpha_{\sigma^2} \to 0$ and $\beta_{\sigma^2} \to 0$, can be used in practice.

$\mathbf{A}$ and σ^2 can also be sampled jointly, i.e, from the joint distribution $p(\mathbf{A}, \sigma^2|\mathbf{s}, \mathbf{x})$. This is done by first sampling $\sigma^{2\,(l)}$ from $p(\sigma^2|\mathbf{s}, \mathbf{x})$ and then sampling $\mathbf{A}^{(l)}$ from $p(\mathbf{A}|\sigma^{2\,(l)}, \mathbf{s}, \mathbf{x})$. The first step involves integrating $\mathbf{A}$ out of the conditional distribution $p(\sigma^2|\mathbf{A}, \mathbf{s}, \mathbf{x})$, which is done in [22].

When the dictionary $\boldsymbol{\Phi}$ is orthonormal and the noise $\mathbf{e}$ is iid Gaussian, $\mathbf{A}$ and σ^2 can be updated using $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{x}}$ instead of $\mathbf{s}$ and $\mathbf{x}$.

2.3.3 Update of the sources

Several strategies can be employed to update the sources $s_1, \ldots, s_n$. The most simple one, and more general, is to update the sources one by one, each conditionally upon the others. Indeed, we have, $\forall t = 1, \ldots, N$,

$$\mathbf{x}_t = s_{i,t}\, \mathbf{a}_i + \sum_{j \neq i} s_{j,t}\, \mathbf{a}_j + \mathbf{e}_t \tag{31}$$

yielding

$$\underbrace{\frac{\mathbf{a}_i^T\, \mathbf{x}}{\mathbf{a}_i^T\, \mathbf{a}_i} - \sum_{j \neq i} \frac{\mathbf{a}_i^T\, \mathbf{a}_j}{\mathbf{a}_i^T\, \mathbf{a}_i}\, s_j}_{x_{i|-i}} = s_i + \underbrace{\frac{\mathbf{a}_i^T\, \mathbf{e}}{\mathbf{a}_i^T\, \mathbf{a}_i}}_{e_{i|-i}} \tag{32}$$

The update of source s_i conditionally upon s_{-i}, σ^2 and $\mathbf{A}$ can thus be recast into a simple unidimensional sparse regression problem with data $x_{i|-i}$ such that

$$x_{i|-i} = \tilde{s}_i\, \boldsymbol{\Phi} + e_{i|-i} \tag{33}$$

and where $e_{i|-i}$ is a Gaussian iid residue of variance $\sigma_i^2 = \sigma^2 / \|\mathbf{a}_i\|_2^2$. Note that, in this framework, the sources can have different models. For example, speech sources and musical sources can be

[3] In practice the columns of $\mathbf{A}$ are normalized to 1 after each draw to solve the BSS indeterminacy on gain. A rigorous implementation would imply sampling each column of $\mathbf{A}$ from the sphere, which is not straightforward. The proposed scheme, though approximate, proved to be satisfactory in practice.

modeled differently, and be processed iteratively. In some cases, when the sources share the same model, block updates of the sources can also be made, as discussed in the following.

When $\mathbf{\Phi}$ is an orthonormal basis, we simply have

$$\tilde{x}_{i|-i} = \tilde{s}_i + \tilde{e}_{i|-i} \tag{34}$$

With the noise and source coefficients iid assumptions, the posterior distribution of $\tilde{s}_i$ factorizes and we simply need to infer $\tilde{s}_{i,k}$ from $\tilde{x}_{i|-i,k}$, $\forall k = 1, \ldots, N$, where $\tilde{s}_{i,k}$ has either a Student t distribution or a hierarchical mixture prior.

2.3.4 Update of Student t source coefficients

We need to sample from $p(\tilde{s}_{i,k}|\alpha_i, \lambda_i, \sigma_i^2, \tilde{x}_{i|-i,k})$. Sampling from this distribution directly is not straightforward, and we instead take advantage of the SMoG formulation of the Student t distribution. As such, a sample $\tilde{s}_{i,k}^{(l)}$ of $p(\tilde{s}_{i,k}|\alpha_i, \lambda_i, \sigma_i^2, \tilde{x}_{i|-i,k})$ can be obtained by sampling $(\tilde{s}_{i,k}^{(l)}, v_{i,k}^{(l)})$ from $p(\tilde{s}_{i,k}, v_{i,k}|\alpha_i, \lambda_i, \sigma_i^2, \tilde{x}_{i|-i,k})$. This is easily done using two Gibbs steps, by alternatively sampling $p(v_{i,k}|\tilde{s}_{i,k}, \alpha_i, \lambda_i)$ and $p(\tilde{s}_{i,k}|v_{i,k}, \sigma_i^2, \tilde{x}_{i|-i,k})$. Conditionally upon $v_{i,k}$, inferring $\tilde{s}_{i,k}$ simply amounts to inferring a Gaussian parameter embedded in Gaussian noise, i.e, Wiener filtering. Thus, we simply have

$$p(\tilde{s}_{i,k}|v_{i,k}, \sigma_i^2, \tilde{x}_{i|-i,k}) = \mathcal{N}(\tilde{s}_{i,k}|\mu_{\tilde{s}_{i,k}}, \sigma^2_{\tilde{s}_{i,k}}) \tag{35}$$

with $\sigma^2_{\tilde{s}_{i,k}} = \left(1/\sigma_i^2 + 1/v_{i,k}\right)^{-1}$ and $\mu_{\tilde{s}_{i,k}} = (\sigma^2_{\tilde{s}_{i,k}}/\sigma_i^2)\, \tilde{x}_{i|-i,k}$.

The sources can alternatively be sampled jointly. Indeed, we have, $\forall k = 1, \ldots, N$

$$\tilde{\mathbf{x}}_k = \mathbf{A}\, \tilde{\mathbf{s}}_k + \tilde{\mathbf{e}}_k \tag{36}$$

Conditionally upon $\mathbf{v}_k = [v_{1,k}, \ldots, v_{n,k}]^T$, the vector $\tilde{\mathbf{s}}_k$ is Gaussian with density $\mathcal{N}(\tilde{\mathbf{s}}_k|0, \mathrm{diag}\,(\mathbf{v}_k))$ and can thus be inferred from $\tilde{\mathbf{x}}_k$ again by Wiener filtering, yielding

$$p(\tilde{\mathbf{s}}_k|\mathbf{A}, \sigma^2, \tilde{\mathbf{x}}_k) = \mathcal{N}(\tilde{\mathbf{s}}_k|\boldsymbol{\mu}_{\tilde{\mathbf{s}}_k}, \mathbf{\Sigma}_{\tilde{\mathbf{s}}_k}) \tag{37}$$

with $\mathbf{\Sigma}_{\tilde{\mathbf{s}}_k} = \left(\frac{1}{\sigma^2}\, \mathbf{A}^T\, \mathbf{A} + \mathrm{diag}\,(\mathbf{v}_k)^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\tilde{\mathbf{s}}_k} = \frac{1}{\sigma^2}\, \mathbf{\Sigma}_{\tilde{\mathbf{s}}_k}\, \mathbf{A}^T\, \tilde{\mathbf{x}}_k$. If block-sampling the sources yields theoretically faster convergence to the stationary distribution, in practice, the two approaches ("one by one" or "full block") involve different computational burdens. With the one by one approach the update of one source $\tilde{s}_i$ requires sampling N univariate independent Gaussian random variables. This does not require any matrix inversion and can be efficiently vectorized in the implementation. The operation has to be repeated n times to update the whole matrix $\tilde{\mathbf{s}}$. On the opposite, with the full block approach, the update of $\tilde{\mathbf{s}}$ requires sampling N times from a n-multivariate Gaussian distribution. This involves inverting a $n \times n$ matrix for each k, and cannot be vectorized, hence requiring to loop over k at each iteration of the Gibbs sampler. The full block approach is thus much more time consuming, and the gain in convergence speed might not be worth the heavier computational burden, as discussed in Section 4.

$v_{i,k}$ has a conjugate prior and its posterior distribution is also inverted-Gamma, such that

$$p(v_{i,k}|\tilde{s}_{i,k}, \alpha_i, \lambda_i) = \mathcal{IG}(v_{i,k}|\alpha_i^{post}, \lambda_{i,k}^{post}) \tag{38}$$

where $\alpha_i^{post} = \frac{1}{2} + \alpha_i$ and $\lambda_{i,k}^{post} = \frac{\tilde{s}_{i,k}^2}{2} + \lambda_i$. Note that here, as opposed to the posterior distribution of σ^2, the influence of the prior in the posterior of $v_{i,k}$ is high, because $v_{i,k}$ is observed through only one data point, $\tilde{s}_{i,k}$.

The posterior distribution of the scale parameter is

$$p(\lambda_i|\alpha_i, v_i) = \mathcal{G}\left(\lambda_i|\alpha_{\lambda_i}^{post}, \beta_{\lambda_i}^{post}\right) \tag{39}$$

Algorithm 2 Gibbs sampler for the Student t prior

Initialize $\boldsymbol{\theta}$
for $l = 1 : L + L_{bi}$ **do**
  Update A and σ^2
  $\mathbf{A} \sim \prod_i \mathcal{N}(\mathbf{r}_i|\boldsymbol{\mu}_{\mathbf{r}_i}, \boldsymbol{\Sigma}_{\mathbf{r}})$
  $\sigma^2 \sim \mathcal{IG}(\sigma^2|\alpha_{\sigma^2}^{post}, \beta_{\sigma^2}^{post})$
  Update source coefficients
  for $i = 1 : n$ **do**
    Update $\tilde{s}_i$
    $\tilde{s}_i \sim \prod_k \mathcal{N}(\tilde{s}_{i,k}|\mu_{\tilde{s}_{i,k}}, \sigma^2_{\tilde{s}_{i,k}})$
    Update v_i
    $v_{i,k} \sim \prod_k \mathcal{IG}(v_{i,k}|\alpha_i^{post}, \lambda_{i,k}^{post})$
    Update λ_i
    $\lambda_i \sim \mathcal{G}\left(\lambda_i|\alpha_{\lambda_i}^{post}, \beta_{\lambda_i}^{post}\right)$
    Update α_i
    $\alpha_i = \text{argmax } p(\alpha_i|\lambda_i, v_i)$
  end for
end for

with $\alpha_{\lambda_i}^{post} = N\,\alpha_i + \alpha_{\lambda_i}$ and $\beta_{\lambda_i}^{post} = \sum_k \frac{1}{v_{i,k}} + \beta_{\lambda_i}$. Again, in this case enough data is available to diminish the influence of the prior, and a Jeffrey noninformative prior $p(\lambda_i) \propto 1/\lambda_i$, corresponding to $\alpha_{\lambda_i} = \beta_{\lambda_i} = 0$ can be used for λ_i.

The posterior distribution of α_i is written

$$p(\alpha_i|\lambda_i, v_i) \propto \exp\left(-N\,\log\Gamma(\alpha_i) + \left(\sum_{k=1}^{N} \log\frac{\lambda_i}{v_{i,k}} - \beta_{\alpha_i}\right)\,\alpha_i\right) \tag{40}$$

This distribution is not straightforward to sample. An exact Metropolis-Hastings (M-H) scheme is described in Appendix (A.2). Alternatively, since the precise value of α_i is unlikely to be important provided it lies within an appropriate small range, this parameter can be sampled from a grid of discrete values with probability mass proportional to (40), like in [22]. Though deterministic moves can compromise the theoretical convergence of the sampler, we also found satisfactory in practice to update α_i to the mode of its posterior distribution. The mode cannot be computed analytically, but can be found using a Newton descent. Table 3 recapitulates the Gibbs sampler for the Student t prior.

2.3.5 Update of source coefficients with the mixture prior

As before, the hierarchical mixture prior does not allow to sample the sources directly from $p(\tilde{s}_{i,k}|\lambda_i, P_i, \sigma_i^2, \tilde{x}_{i|-i,k})$ but rather requires sampling jointly from $p(\tilde{s}_{i,k}, \gamma_{i,k}, v_{i,k}|\lambda_i, P_i, \sigma_i^2, \tilde{x}_{i|-i,k})$. Again, this done by Gibbs sampling, more precisely, alternate sampling from $p(\tilde{s}_{i,k}, \gamma_{i,k}|v_{i,k}, P_i, \sigma_i^2, \tilde{x}_{i|-i,k})$ and $p(v_{i,k}|\tilde{s}_{i,k}, \gamma_{i,k}, \lambda_i)$. The first distribution can be sampled directly, by

1) sampling $\gamma_{i,k}^{(l)}$ from $P(\gamma_{i,k}|v_{i,k}, P_i, \sigma_i^2, \tilde{x}_{i|-i,k})$

2) sampling $\tilde{s}_{i,k}^{(l)}$ from $p(\tilde{s}_{i,k}|\gamma_{i,k}^{(l)}, v_{i,k}, \sigma_i^2, \tilde{x}_{i|-i,k})$

Note that these two latter steps are not Gibbs sampling steps, but constitute an exact draw from the joint distribution.[4] In fact, as pointed out in [23], a Gibbs implementation consisting of sampling alternatively $\tilde{s}_{i,k}|\gamma_{i,k}$ and $\gamma_{i,k}|\tilde{s}_{i,k}$ cannot be used as it leads to a nonconvergent Markov

[4]To make things perfectly clear, let us recall that $(x^{(l)}, y^{(l)})$ can be drawn from $p(x, y)$ in two different ways. An exact draw is obtained by sampling $x^{(l)}$ from $p(x)$ and then $y^{(l)}$ from $p(y|x^{(l)})$ (or the other way round). This requires to be able to sample from $p(x)$ directly. If this not possible, the other way is to do Gibbs sampling, i.e sampling alternatively $p(x|y)$ and $p(y|x)$, until the stationary distribution is obtained.

chain (the Gibbs sampler gets stuck when it generates a value $\tilde{s}_{i,k} = 0$). We thus need to evaluate $P(\gamma_{i,k}|v_{i,k}, P_i, \sigma_i^2, \tilde{x}_{i|-i,k})$. This can be seen as an hypothesis testing problem, with

$$(H_1) \iff \gamma_{i,k} = 1 \iff \tilde{x}_{i|-i,k} = \tilde{s}_{i,k} + \tilde{e}_{i|-i,k} \quad (41)$$

$$(H_0) \iff \gamma_{i,k} = 0 \iff \tilde{x}_{i|-i,k} = \tilde{e}_{i|-i,k} \quad (42)$$

The ratio

$$\tau_{i,k}^{post} = \frac{P(\gamma_{i,k} = 1|v_{i,k}, P_i, \sigma_i^2, \tilde{x}_{i|-i,k})}{P(\gamma_{i,k} = 0|v_{i,k}, P_i, \sigma_i^2, \tilde{x}_{i|-i,k})}$$

is thus simply expressed as

$$\tau_{i,k}^{post} = \frac{\mathcal{N}(\tilde{x}_{i|-i,k}|0, v_{i,k} + \sigma_i^2)}{\mathcal{N}(\tilde{x}_{i|-i,k}|0, \sigma_i^2)} \frac{P_i}{1 - P_i} \quad (43)$$

$$= \sqrt{\frac{\sigma^2}{\sigma^2 + v_{i,k}}} \exp\left(\frac{\tilde{x}_{i|-i,k}^2 \, v_{i,k}}{2\,\sigma^2(\sigma^2 + v_{i,k})}\right) \frac{P_i}{1 - P_i} \quad (44)$$

$\gamma_{i,k}$ is thus drawn from the two states discrete distribution with probability masses

$$P(\gamma_{i,k} = 0|v_{i,k}, P_i, \sigma_i^2, \tilde{x}_{i|-i,k}) = 1/(1 + \tau_{i,k}^{post}) \quad (45)$$

$$P(\gamma_{i,k} = 1|v_{i,k}, P_i, \sigma_i^2, \tilde{x}_{i|-i,k}) = \tau_{i,k}^{post}/(1 + \tau_{i,k}^{post}) \quad (46)$$

When a value $\gamma_{i,k} = 0$ is drawn, $\tilde{s}_{i,k}$ is set to zero. Otherwise it is updated through Wiener filtering as before. The posterior distribution of $\tilde{s}_{i,k}$ is thus written as

$$p(\tilde{s}_{i,k}|\gamma_{i,k}, v_{i,k}, \sigma^2, \tilde{x}_{i|-i,k}) = (1 - \gamma_{i,k})\,\delta_0(\tilde{s}_{i,k}) + \gamma_{i,k}\,\mathcal{N}(\tilde{s}_{i,k}|\mu_{\tilde{s}_{i,k}}, \sigma^2_{\tilde{s}_{i,k}}) \quad (47)$$

with $\sigma^2_{\tilde{s}_{i,k}}$ and $\mu_{\tilde{s}_{i,k}}$ defined as before.

It is also possible to make block draws of $(\tilde{\mathbf{s}}_k, \boldsymbol{\gamma}_k)$, where $\boldsymbol{\gamma}_k = [\gamma_{1,k}, \ldots, \gamma_{n,k}]^T$. This strategy amounts to solving for every iteration l and every index k a small sized Bayesian variable selection problem, where $\mathbf{A}$ contains the regressors and $\tilde{\mathbf{s}}_k$ contains the regression coefficients [23]. It requires

1) sampling $\boldsymbol{\gamma}_k$ from $p(\boldsymbol{\gamma}_k|\mathbf{v}_k, \boldsymbol{\theta}_\gamma, \sigma^2, \tilde{\mathbf{x}})$, which requires computing 2^n probability masses corresponding to the 2^n values of $\boldsymbol{\gamma}_k$,

2) sampling $\tilde{\mathbf{s}}_k$ from $p(\tilde{\mathbf{s}}_k|\boldsymbol{\gamma}_k, \mathbf{v}_k, \sigma^2, \tilde{\mathbf{x}})$ which is multivariate Gaussian,

where $\boldsymbol{\theta}_\gamma$ contains all the Bernoulli probabilities. The computation of $p(\boldsymbol{\gamma}_k|\mathbf{v}_k, \boldsymbol{\theta}_\gamma, \sigma^2, \tilde{\mathbf{x}})$ involves the integration of $\mathcal{N}(\tilde{\mathbf{x}}_k|\mathbf{A}\,\tilde{\mathbf{s}}_k, \sigma^2\,\mathbf{I}_m)\,\mathcal{N}(\tilde{\mathbf{s}}_k, |0, \mathrm{diag}\,(\mathbf{v}_k))$ over $\tilde{\mathbf{s}}$, which is analytically possible, but involves small size matrix inversions for each value of $\boldsymbol{\gamma}_k$. The full block approach is too computationally demanding, and is still inapplicable in practice.

The conditional posterior distribution of $v_{i,k}$ is

$$p(v_{i,k}|\gamma_{i,k}, \tilde{s}_{i,k}, \lambda_i) = (1 - \gamma_{i,k})\,\mathcal{IG}\,(v_{i,k}|\alpha_i, \lambda_i) + \gamma_{i,k}\,\mathcal{IG}\left(v_{i,k}|\alpha_i^{post}, \lambda_{i,k}^{post}\right) \quad (48)$$

When a value $\gamma_{i,k} = 0$ is generated, $\tilde{v}_{i,k}$ is simply sampled from its prior (no posterior information is available), otherwise, it is inferred from the available value of $\tilde{s}_{i,k}$ as before.

The posterior distribution of the scale parameter λ_i is unchanged, and as given by (39). However, because we are looking for sparse representations, most of the indicator variables $\gamma_{i,k}$ take the value 0 and thus most of the variances $v_{i,k}$ are sampled from their prior (see (48)). Thus, the

Algorithm 3 Gibbs sampler for the hierarchical prior

Initialize $\boldsymbol{\theta}$
for $l = 1 : L + L_{bi}$ **do**
  Update A and σ^2
  $\mathbf{A} \sim \prod_i \mathcal{N}(\mathbf{r}_i|\boldsymbol{\mu}_{\mathbf{r}_i}, \boldsymbol{\Sigma}_{\mathbf{r}})$
  $\sigma^2 \sim \mathcal{IG}(\sigma^2|\alpha_{\sigma^2}^{post}, \beta_{\sigma^2}^{post})$
  Update sources
  for $i = 1 : n$ **do**
    Update γ_i **and** $\tilde{s}_i$
    $\gamma_i \sim \prod_k P(\gamma_{i,k}|\tau_{i,k}^{post})$
    $\tilde{s}_i \sim \prod_k (1 - \gamma_{i,k})\, \delta_0(\tilde{s}_{i,k}) + \gamma_{i,k}\, \mathcal{N}(\tilde{s}_{i,k}|\mu_{\tilde{s}_{i,k}}, \sigma^2_{\tilde{s}_{i,k}})$
    Update $\{v_{i,k} : \gamma_{i,k} = 1\}$
    $v_{i,k} \sim \mathcal{IG}(v_{i,k}|\alpha_i^{post}, \lambda_{i,k}^{post})$
    Update λ_i
    $\lambda_i \sim \mathcal{G}\left(\lambda_i|\alpha_{\lambda_i}^{post'}, \beta_{\lambda_i}^{post'}\right)$
    Update $\{v_{i,k} : \gamma_{i,k} = 0\}$
    $v_{i,k} \sim \mathcal{IG}(v_{i,k}|\alpha_i, \lambda_i)$
    Update P_i
    $P_i \sim \mathcal{B}(P_i|\alpha_{P_i}^{post}, \beta_{P_i}^{post})$
  end for
end for

influence of the data in the full posterior distribution of λ_i becomes small, and the convergence of λ_i can be very slow. A faster scheme, employed in [24, 25], consists of making one draw from $p(\{v_{i,k} : \gamma_{i,k} = 1\}|\{\tilde{s}_{i,k} : \gamma_{i,k} = 1\}, \lambda_i)$, then one draw from $p(\lambda_i|\{v_{i,k} : \gamma_{i,k} = 1\})$ and finally one draw from $p(\{v_{i,k} : \gamma_{i,k} = 0\}|\lambda_i)$. The posterior $p(\lambda_i|\{v_{i,k} : \gamma_{i,k} = 1\})$ is simply written

$$p(\lambda_i|\alpha_i, v_i) = \mathcal{G}\left(\lambda_i|\alpha_{\lambda_i}^{post'}, \beta_{\lambda_i}^{post'}\right) \tag{49}$$

with $\alpha_{\lambda_i}^{post'} = \#\gamma_i\, \alpha_i + \alpha_{\lambda_i}$ and $\beta_{\lambda_i}^{post'} = \sum_{k:\gamma_{i,k}=1} 1/v_{i,k} + \beta_{\lambda_i}$, and where $\#\gamma_i$ is the number of values in γ_i equal to 1.

Finally, the posterior distribution of P_i is

$$p(P_i|\gamma_i) = \mathcal{B}(P_i|\alpha_{P_i}^{post}, \beta_{P_i}^{post}) \tag{50}$$

where $\alpha_{P_i}^{post} = \#\gamma_i + \alpha_{P_i}$, $\beta_{P_i}^{post} = N - \#\gamma_i + \beta_{P_i}$.

Table 3 recapitulates the Gibbs sampler for the hierarchical prior.

3 Audio-specific models

While the previous section described models for sparse sources in general, in this section we focus on audio source separation. As such we bring modifications to the general priors described previously in order to take into account the specificities of audio signals. In particular we lift the iid assumption on $\tilde{s}_i$. In Section 3.2, the "identically distributed" modeling is replaced by frequency-dependent priors which model the natural non-uniform distribution of energy along frequency. In Section 3.3, the "independently distributed" modeling is replaced by structured priors which model harmonic properties of sound (persistence of time-frequency coefficients through time). The audio sources are here modeled as a sparse linear combination of Modified Discrete Cosine Transform (MDCT) atoms which is a time-frequency orthonormal basis presented in next the section.

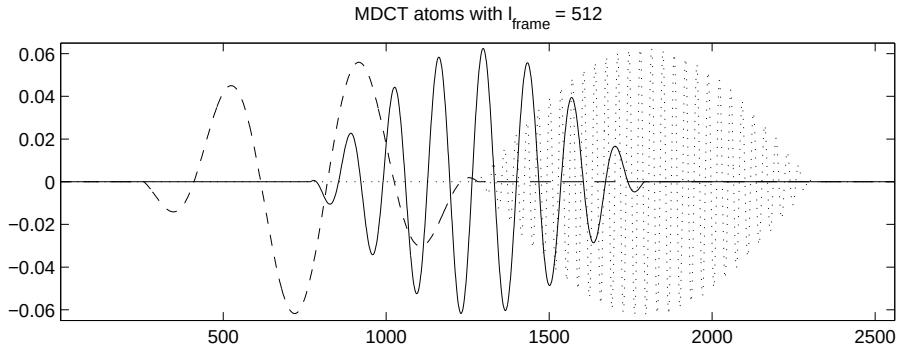

Figure 4: Three MDCT atoms for $l_{\text{frame}} = 512$ and using a sine bell window. The atoms are adjacent in time (50 % overlap) and correspond to three different value of frequency q.

3.1 MDCT representation

The MDCT basis is a local cosine lapped transform [14], which has proven to give good sparse approximations of audio signals, with many coding applications [26, 27]. It provides an orthonormal decomposition without blocking effects, and has fast implementations based on the FFT. Atoms corresponding to the MDCT transform of a signal of length $N = l_{\text{frame}} \times n_{\text{frame}}$ and a frame length l_{frame}, are defined as

$$\boldsymbol{\Phi}_{(q,p)}(t) = w(t-(p-1)\,l_{\text{frame}})\,\cos\!\left[\frac{\pi}{l_{\text{frame}}}\left(t-(p-1)\,l_{\text{frame}}+\frac{l_{\text{frame}}+1}{2}\right)\left(q-\frac{1}{2}\right)\right] \quad (51)$$

where $w(t)$ is a window of size $2\,l_{\text{frame}}$, $q = 1, \ldots, l_{\text{frame}}$ is a frequency index and $p = 1, \ldots, n_{\text{frame}}$ is a time frame index. When convenient, we will use in the following the mapping $k = (p-1)\,l_{\text{frame}}+q$ to index the atoms of the dictionary $\boldsymbol{\Phi}$. Fig. 4 shows several instances of MDCT atoms.

3.2 Frequency-dependent models

The energy distribution of audio signal is naturally decreasing across frequency. Thus, an identically distributed model for $\tilde{s}_i$ does not best fit audio data. However, the models presented in Section 2.1 can easily be made frequency dependent by weighting the scale parameter λ_i with a frequency profile. As such, the prior about $v_{i,k}$ becomes

$$p(v_{i,k}|\alpha_i, \lambda_i) = \mathcal{IG}(v_{i,k}|\alpha_i, \lambda_i\, f_k) \quad (52)$$

with

$$f_k = f_q = \frac{1}{(1+((q-1)/q_c)^R)}, \quad q = 1, \ldots, l_{\text{frame}} \quad (53)$$

This frequency shaping is based on the frequency response of a Butterworth lowpass filter, where q_c acts as a cut-off frequency and R acts as the filter order. In practice we set $R = 2$, $q_c = l_{\text{frame}}/3$.

3.3 Structural constraints

The harmonic nature of sound generates spectral lines in the time-frequency plane. The frame length of the MDCT basis being typically 22 ms, a sound event (e.g, a partial of one note) usually lies over several adjacent atoms. Thus, when an atom is selected at frame p and frequency q, it is "likely" that atoms at frame $p-1$ and $p+1$ are selected too, at the same frequency q.[5] This property is not described by the "independently distributed" models of Section 2.1. Instead,

[5] In fact, it could be at frequency $q-1$, q or $q+1$, but this more general case is not considered here, though it readily fits in the framework we describe.

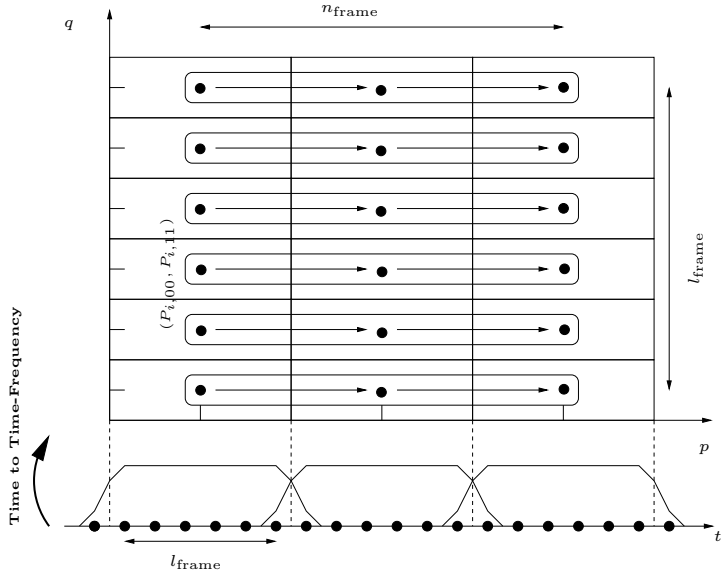

Figure 5: This figure illustrates the structured hierarchical mixture prior. Each square of the time-frequency tiling corresponds to a MDCT atom. To each atom corresponds an indicator variable $\gamma_{i,k}$ which controls whether this atom is selected ($\gamma_{i,k} = 1$) or not ($\gamma_{i,k} = 0$). The set of indicator variables γ_i is modeled as "horizontal and parallel" Markov chains of order 1, with common transition probabilities $P_{i,00}$ and $P_{i,11}$, and with initial probability taken as its equilibrium value.

source coefficients should be correlated through time in order to favor horizontal structures in the time-frequency plane. This can be done easily with the mixture prior by replacing the independent Bernoulli prior for γ_i with a *structured prior*, such as Markov chain modeling. As such, for a fixed frequency index q the sequence $\{\gamma_{i,(q,p)}\}_{p=1,\ldots,n_{\text{frame}}}$ is modeled by a 2-state first order Markov chain with transition probabilities $P_{i,00}$ and $P_{i,11}$, assumed equal for all frequency indices $q = 1, \ldots, l_{\text{frame}}$. The initial distribution $\pi_i = P(\gamma_{i,(q,1)} = 1)$ of each chain is taken to be its stationary distribution in order to impose some form of shift-invariance [29], namely

$$\pi_i = \frac{1 - P_{i,00}}{2 - P_{i,11} - P_{i,00}} \quad \text{and} \quad (1 - \pi_i) = \frac{1 - P_{i,11}}{2 - P_{i,11} - P_{i,00}} \tag{54}$$

This type of model was introduced for speech denoising in [24]. The model is illustrated in Fig. 5, and the corresponding graph for $\tilde{s}_{i,k}$ is given in Fig. 6

The Markov transition probabilities are estimated and given Beta priors $\mathcal{B}(P_{i,00}|\alpha_{P_{i,00}}, \beta_{P_{i,00}})$ and $\mathcal{B}(P_{i,11}|\alpha_{P_{i,11}}, \beta_{P_{i,11}})$.

3.4 Modified inference steps

The use of a weight function with λ_i does not change much in the Gibbs sampler. Basically, the weight function contribution must be mirrored everywhere λ_i appears. As such, the parameters of the posterior distributions of $v_{i,k}$ and λ_i are solely changed as

$$\lambda_{i,k}^{post} = \frac{\tilde{s}_{i,k}^2}{2} + \lambda_i \, f_k \tag{55}$$

and

$$\beta_{\lambda_i}^{post} = \sum_k \frac{f_k}{v_{i,k}} + \beta_{\lambda_i} \tag{56}$$

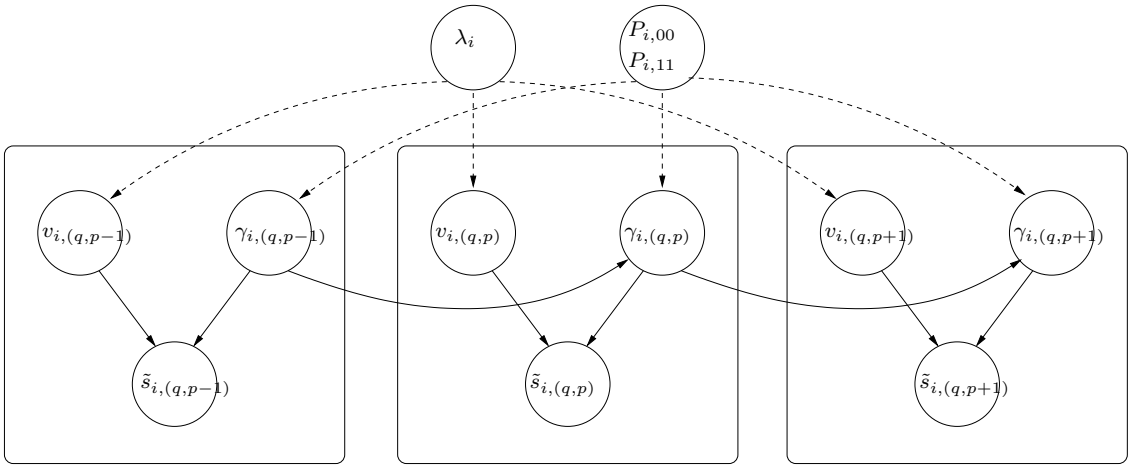

Figure 6: Graph model of the source coefficients with the structured hierarchical mixture prior.

Furthermore, the posterior distribution of α_i writes

$$p(\alpha_i|\lambda_i, v_i) \propto \exp\left(-N\ \log\Gamma(\alpha_i) + \left(\sum_{k=1}^{N} \log\frac{\lambda_i\, f_k}{v_{i,k}} - \beta_{\alpha_i}\right)\alpha_i\right) \tag{57}$$

and its mode can again be computed using Newton descent rooting.

The update of the source coefficients with the structured hierarchical mixture prior is not much more difficult. As before, starting from (33), which writes

$$\tilde{x}_{i|-i} = \tilde{s}_i + \tilde{e}_{i|-i}$$

we wish to sample from $p(\tilde{s}_i, \gamma_i|v_i, \sigma_i^2, \tilde{x}_{i|-i}) = p(\tilde{s}_i|\gamma_i, v_i, \sigma_i^2, \tilde{x}_{i|-i})\, P(\gamma_i|v_i, \sigma_i^2, \tilde{x}_{i|-i})$. Conditionally upon $\gamma_{i,k}$, $\gamma_{i,k'}$, the coefficients $\tilde{s}_{i,k}$, $\tilde{s}_{i,k'}$ are still mutually independent, correlation being introduced through the indicator variables only, see Fig. 6. Thus, the distribution $p(\tilde{s}_i|\gamma_i, v_i, \sigma_i^2, \tilde{x}_{i|-i})$ still factorizes as $p(\tilde{s}_i|\gamma_i, v_i, \sigma_i^2, \tilde{x}_{i|-i}) = \prod_k p(\tilde{s}_{i,k}|\gamma_{i,k}, v_{i,k}, \sigma_i^2, \tilde{x}_{i|-i,k})$. The distribution $P(\gamma_i|v_i, \sigma_i^2, \tilde{x}_{i|-i})$ however, does not factorize anymore. Instead, we have to condition the update of $\gamma_{i,k}$ upon the set $\gamma_{i,-k}$, such that

$$\tau_{i,k}^{post} = \frac{P(\gamma_{i,k}=1|\gamma_{i,-k}, v_{i,k}, \sigma_i^2, \tilde{x}_{i|-i,k})}{P(\gamma_{i,k}=0|\gamma_{i,-k}, v_{i,k}, \sigma_i^2, \tilde{x}_{i|-i,k})} \tag{58}$$

$$= \frac{\mathcal{N}(\tilde{x}_{i|-i,k}|0, v_{i,k}+\sigma_i^2)}{\mathcal{N}(\tilde{x}_{i|-i,k}|0, \sigma_i^2)}\, \frac{P(\gamma_{i,k}=1|\gamma_{i,-k}, \theta_{\gamma_i})}{P(\gamma_{i,k}=0|\gamma_{i,-k}, \theta_{\gamma_i})} \tag{59}$$

This expression of $\tau_{i,k}^{post}$ only differs from the one given by (44) in the ratio

$$\tau_{i,k} = P(\gamma_{i,k}=1|\gamma_{i,-k}, \theta_{\gamma_i})/P(\gamma_{i,k}=0|\gamma_{i,-k}, \theta_{\gamma_i}),$$

which is given in Appendix A.3. Note that the chosen structure for the coefficients is solely reflected in the inference part through $\tau_{i,k}$. If any other structure was to be considered, such as Markov random fields (like in [24]), one would only need to change $\tau_{i,k}$ accordingly. All the other update steps remain unchanged.

Because we have assumed the initial probability of the chain to be equal to its equilibrium probability, the posterior distributions of $P_{i,00}$ and $P_{i,11}$ do not belong to a family of distributions easy to sample. Their expressions are given in Appendix A.4 where we describe an exact M-H scheme as well as a deterministic scheme to update these variables.

4 Separation of a stereophonic recording with 3 sources

4.1 Experimental setup

We present results for blind separation of a stereo mixture ($m = 2$) of $n = 3$ musical sources (voice, acoustic guitar, bass guitar).[6] The sources were obtained from the BASS-dB database [31]. They consist of excerpts of original tracks from the song *Anabelle Lee* (Alex Q), published under a Creative Commons Licence. The signals are sampled at $f_s = 22.5kHz$ with length $N = 131072$ ($\approx 6s$). The mixing matrix is given in Table 1; it provides a mixture where the voice s_1 is in the middle, the acoustic guitar s_2 originates at 67.5^o on the left and the bass guitar s_3 at 67.5^o on the right. Gaussian noise was added to the observations with $\sigma = 0.01$, resulting in respectively $25dB$ and $27dB$ input SNR on each channel. We applied a MDCT to the observations using a sine bell and 50% overlap, with time resolution (half the window length) $l_{\text{frame}} = 512$ (22ms). We compare the following methods:

(a) The source coefficients are given a Student t iid distribution. The sources are updated with block draws of $\tilde{\mathbf{s}}_k$. Using a MATLAB implementation running on a 1.25GHz Powerbook G4 with 512 MB RAM, 1000 iterations of the sampler take 6.6 hours.

(b) We apply the approach (a), but the sources are updated one by one, conditionally upon the others. 1000 iterations of the sampler take 1.1 hours.

(c) Same as (b) except that the source coefficients now have a hierarchical mixture frequency-dependent prior with Bernoulli prior on γ_i. 1000 iterations of the sampler take 50 min.

(d) Same as (c) except that the source coefficient now have the structured prior described in Section 3.3 (horizontal Markov chain modeling of γ_i). The computational burden is nearly unchanged, 1000 iterations take approximately 50min.

Hyperparameters priors where all chosen so as to yield noninformative priors, i.e, $\alpha_{\lambda_i} = \beta_{\lambda_i} = 0$, $\beta_{\alpha_i} = 0$, $\alpha_{P_i} = \beta_{P_i} = 0$, $\alpha_{P_{i,00}} = \beta_{P_{i,00}} = \alpha_{P_{i,11}} = \beta_{P_{i,11}} = 0$. A Jeffreys noninformative prior was also chosen for σ^2, setting $\alpha_{\sigma^2} = \beta_{\sigma^2} = 0$. $\mathbf{A}$ was initialized to [1 1 1; 0 0 0], $\tilde{s}_i$ to $\tilde{x}_1/3$, v_i to ones, λ_i to 0.1, α_i to 0.5, P_i to 0.1, and the Markov transition probabilities to 0.9.

The samplers were run for 2500 iterations in case (a) and for 10000 iterations in the other cases. Approximate convergence was *generally* observed after 1500 iterations in the first case and after 5000 iterations in the second one. In every case σ^2 was annealed to its true posterior distribution during the first hundreds iterations of the sampler. In our framework, simulated annealing basically consists in artificially increasing the values taken by σ^2 through the first iterations of the sampler and gradually decreasing them to their correct expected value. Increasing σ^2 broadens the support of the likelihood and allows the sampler to explore faster the space of parameters. This was empirically proven to accelerate the convergence of the sampler to its stationary distribution in [22]. We anneal the degrees of freedom α_{σ^2} of the input noise by replacing $\alpha_{\sigma^2}^{post}$ in (30) with $\alpha_{\sigma^2}^{post'}(l) = (1 - (1 - p_0)\exp(-l/l_0))\,\alpha_{\sigma^2}^{post}$, where l 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 $l >> l_0$.

MMSE estimates of the source coefficients were computed in each case by averaging the last 1000 samples. Table 1 gives the mixing matrix estimates as well as separation evaluation criteria for the estimated sources in each case. The criteria are described in [32], but basically, the SDR (Source to Distortion Ratio) provides an overall separation performance criterion, the SIR (Source to Interference Ratio) measures the level of interferences from the other sources in each source estimate, SNR (Source to Noise Ratio) measures the error due to the additive noise on the sensors and the SAR (Source to Artifacts Ratio) measures the level of artifacts in the source estimates. Source estimates can be listened to at [33], which is perhaps the best way to assess the quality of the results.

[6] Part of these results are reproduced from [30].

Original matrix

$$\mathbf{A} = \begin{bmatrix} 0.7071 & 0.9808 & 0.1951 \\ 0.7071 & 0.1951 & 0.98079 \end{bmatrix}$$

Method (a)

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.7074 & 0.9811 & 0.1947 \\ (\pm 0.0003) & (\pm 0.0002) & (\pm 0.0004) \\ 0.7067 & 0.1930 & 0.98085 \\ (\pm 0.0003) & (\pm 0.0009) & (\pm 0.00008) \end{bmatrix}$$

Method (b)

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.7074 & 0.9809 & 0.1948 \\ (\pm 0.0004) & (\pm 0.0002) & (\pm 0.0004) \\ 0.7068 & 0.1944 & 0.98084 \\ (\pm 0.0004) & (\pm 0.0008) & (\pm 0.00007) \end{bmatrix}$$

Method (c)

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.7044 & 0.9821 & 0.1943 \\ (\pm 0.0004) & (\pm 0.0002) & (\pm 0.0006) \\ 0.7098 & 0.1881 & 0.9809 \\ (\pm 0.0004) & (\pm 0.0012) & (\pm 0.0001) \end{bmatrix}$$

Method (d)

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.7079 & 0.9811 & 0.1946 \\ (\pm 0.0003) & (\pm 0.0001) & (\pm 0.0006) \\ 0.7064 & 0.1933 & 0.9809 \\ (\pm 0.0003) & (\pm 0.0007) & (\pm 0.0001) \end{bmatrix}$$

Method	$\hat{s}_1$ (voice)			
	SDR	SIR	SAR	SNR
(a)	4.0	14.1	4.6	28.6
(b)	-2.0	14.0	-1.7	22.3
(c)	0.1	5.6	2.6	28.4
(d)	-0.75	12.0	-0.23	28.3

Method	$\hat{s}_2$ (acoustic guitar)			
	SDR	SIR	SAR	SNR
(a)	5.6	17.7	6.0	28.0
(b)	3.2	6.5	6.9	29.3
(c)	0.7	7.7	2.4	27.4
(d)	3.1	10.3	4.5	38.6

Method	$\hat{s}_3$ (bass guitar)			
	SDR	SIR	SAR	SNR
(a)	10.5	20.9	11.0	40.0
(b)	8.0	11.2	11.1	41.6
(c)	5.9	14.4	6.7	51.5
(d)	7.4	15.1	8.3	50.1

Table 1: Mixing matrix estimates and performance criteria.

Fig. 7 and Fig. 9 respectively represents the sampled values of all the parameters with approach (b) and approach (d). Fig. 8 represents the histograms of the coefficients of the *original* sources compared with the Student t densities estimated with method (b). Fig. 10 presents the *significance maps* of the source coefficients, i.e, the MAP estimates of γ_1, γ_2 and γ_3, in the Bernoulli (method (c)) and Markov (method (d)) cases.

4.2 Discussion

4.2.1 Separation quality

The sound samples show that best perceptual results are obtained with method (d), that is, when audio-specific structured priors are used. The source estimates with this approach may not be the best in terms of interference rejection, but what prevails is that they sound the *cleanest*. Artifacts remain but may be considered less disturbing than in the other sound samples, originating from isolated time-frequency atoms scattered over the time-frequency plane (as illustrated by Fig. 10). These perceptual considerations may not be reflected in the numerical criteria given at Table 1, which only compares the waveforms themselves. In every case the mixing matrix is very well estimated, with very small error standard deviations.

4.2.2 Strengths and limits of the MCMC approach

One striking fact about the numerical criteria given in Table 1 is the difference in results between methods (a) and (b). Though the two methods should theoretically yield similar source estimates after a "large enough" number of iterations, in practice, over an horizon of 10000 iterations, method (a) still yields better estimates, in particular in terms of SIRs. We believe this is because the individual update of each source conditionally upon the others creates some correlation between the sources. If the amount of correlation should theoretically fade away when averaging a large number of samples, well after the burn in period, in practice this seems to be a problem over our limited horizon.[7] We also noticed that, depending on the initializations and the random sequence seeds, method (b), and also (c), could get stuck for long periods in some local modes of the posterior distribution of the mixing matrix, and that full exploration of the posterior could be tedious. This seemed to be a lesser problem with (d), probably because the structure brought in the model regularizes the posterior of the parameters. In contrast, method (a) happened to be more robust to local modes, and convergence is rather fast (in number of iterations) when σ^2 is annealed.

[7] Actually, Fig. 7 shows that after 10000 iterations, some source hyperparameters seem not to have fully converged to their stationary distribution, when this was not the case with method (a), not shown here.

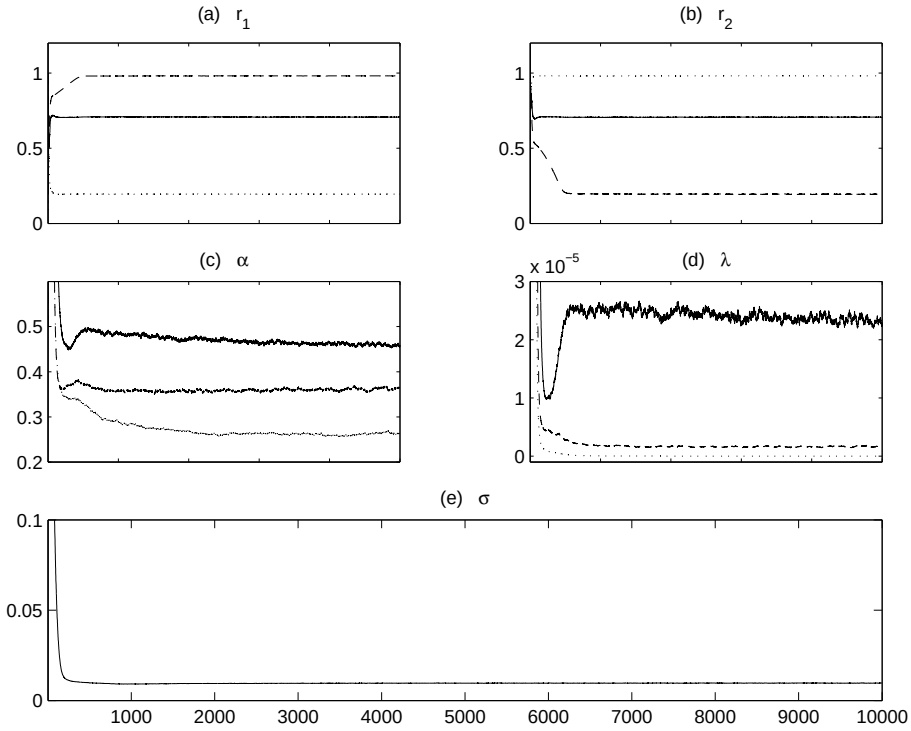

Figure 7: Gibbs samples of the model parameters in the Student t iid case with one by one update of the sources (method (b)); (-): source 1, (- -): source 2, (:) = source 3.

Note however that in every case, the Gibbs sampler fail at exploring *all* the posterior space. Indeed, indeterminacies remain in the model, namely sign and permutation ambiguities between $\mathbf{A}$ and $\mathbf{s}$, corresponding to equivalent modes of the posterior. The Gibbs sampler should explore these modes too, which in practice, over a limited horizon of iterations, was never the case. The sampler converges to one mode, explore this mode, and gets stuck there.

As such, a rigorous use of MCMC methods for practical applications should involve thorough monitoring of the sampler, for example by running several chains with different initializations and comparing the behaviors of each chain [19].

MCMC techniques are nevertheless more robust than EM-like counterparts, though at the expense of higher computation times. They provide a full description of the posterior space, as opposed to point estimates only. Hence, they can be used to compute interval estimates, and also to compute MAP estimates for some parameters and MMSE estimates for others. If MCMC techniques can be considered not yet applicable in practice (but is likely to change owing to the fast evolution of computing facilities), they can at least be used as a diagnostic tool to check the validity of a given source model. Once the model is chosen, lighter optimization techniques can be devised for that particular model.

5 Conclusions

In conclusion, let us emphasize that the Bayesian approach provides a framework to gather all information available about the sources into a prior distribution $p(\tilde{s}_i|\boldsymbol{\theta}_{s_i})$. In the case of linear instantaneous mixtures, and more generally convolutive mixtures, each source can have a different model and be updated separately. As mentioned in Section 2.3.3, more complex models, involving overcomplete dictionaries can even be considered. As such, musical signals can be modeled as a linear combination of MDCT atoms with short time resolution aiming at representing the transient parts, such the attacks of notes and MDCT atoms with longer time resolution (and finer frequency resolution) aiming at representing tonals (like in this chapter) [29]. As well as models of the

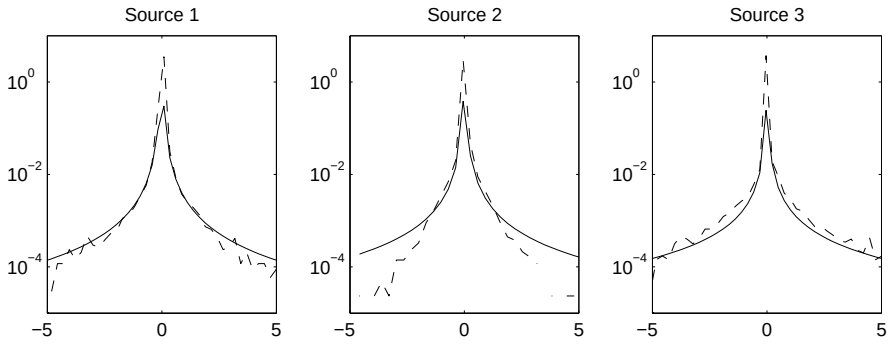

Figure 8: Histograms of the original source coefficients $\tilde{s}_i$ compared to the estimated Student t densities with method (b). The MMSE estimate for α and λ are respectively [0.91, 0.72, 0.52] and [0.0071, 0.0021, 0.0002]. These are close to the ones obtained with method (a), which are respectively [0.84, 0.76, 0.65] and [0.0066, 0.0023, 0.0004].

sources, other models of mixing systems can be designed. Frequency dependent mixing matrices could be used to model convolution effects, while source models could help solving the frequency permutation problem.

Finally, note that, on a wider scope, these models of sources and mixing systems, generative in essence, propose a semantic, object-based, representation of multichannel sound. Their effectiveness is here evaluated for the source separation problem, but they could be employed for other tasks. As such, the sparsity of the maps shown on Fig. 10 suggests efficient coding schemes [34], where each source could be treated separately instead of encoding the mixture directly. Because of the synthesis approach, the sound mixture can also be modified, for example denoised (by simply remixing the estimated sources and mixing system), but also remixed (by changing the parameters of the mixing system), or reshaped (false notes can be corrected by moving or removing the corresponding time-frequency atoms).

A Appendices

A.1 Standard distributions

Multivariate Gaussian	$\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \|2\pi\,\boldsymbol{\Sigma}\|^{-\frac{1}{2}}\,\exp -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})$
Gaussian	$\mathcal{N}(x\|u,\sigma^2) = (2\pi\sigma^2)^{-1/2}\,\exp -\frac{(x-u)^2}{2\sigma^2}$
Beta	$\mathcal{B}(x\|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}\,(1-x)^{\beta-1}, \quad x\in[0,1]$
Gamma	$\mathcal{G}(x\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,x^{\alpha-1}\,\exp(-\beta\,x),\ x\geq 0$
inv-Gamma	$\mathcal{IG}(x\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,x^{-(\alpha+1)}\,\exp(-\frac{\beta}{x}),\ x\geq 0$

The inverted-Gamma distribution is the distribution of $1/X$ when X is Gamma distributed.

A.2 M-H update of the degrees of freedom parameter α_i

α_i can be updated using independent M-H sampling,[8] as proposed in [35]. The posterior distribution of α_i, given by (40) is written

$$p(\alpha_i|\lambda_i, v_i) \propto g(\alpha_i)^N \tag{60}$$

with

$$g(\alpha_i) = \frac{1}{\Gamma(\alpha_i)}\,\exp(-\beta_{\alpha_i}^{post}\,\alpha_i) \tag{61}$$

[8]Independent Metropolis-Hasting sampling is another MCMC technique. Given a proposal distribution $q(x)$, at iteration $l+1$, a candidate $x^\star$ is generated from $q(x)$ and accepted with probability $\min\left\{1, \frac{p(x^\star)\,q(x^{(l)})}{q(x^\star)\,p(x^{(l)})}\right\}$. If accepted $x^{(l+1)} = x^\star$, otherwise $x^{(l+1)} = x^{(l)}$.

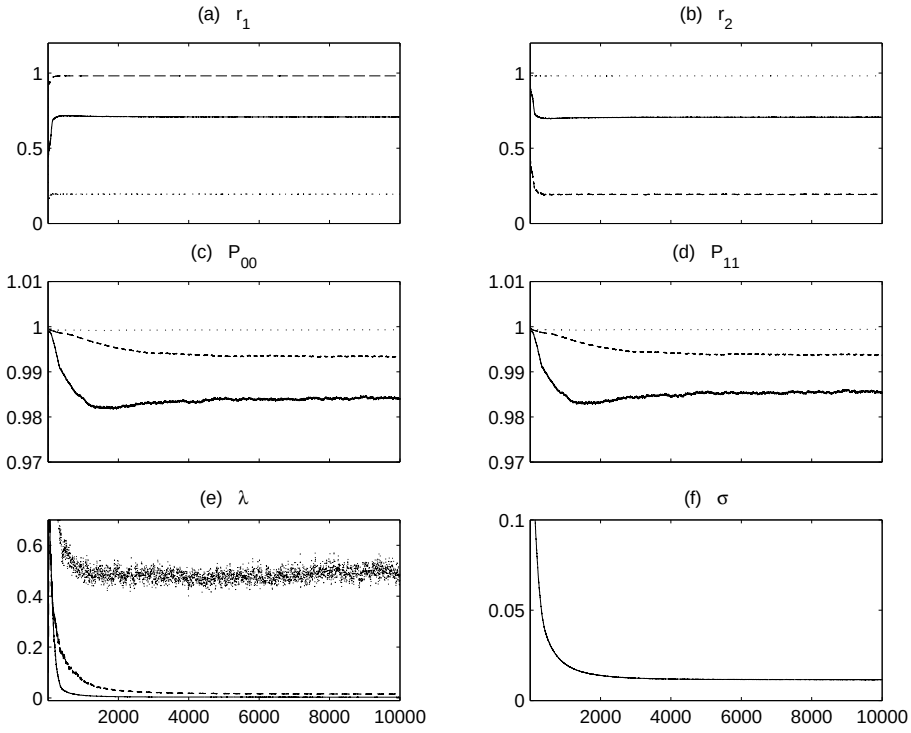

Figure 9: Gibbs samples of the model parameters with the hierarchical frequency dependent mixture prior, with Markov horizontal structure (method (d)); (-): source 1, (- -): source 2, (:) = source 3.

and

$$\beta_{\alpha_i}^{post} = \frac{1}{N}\left(-\sum_{k=1}^{N}\log\frac{\lambda_i}{v_{i,k}} + \beta_{\alpha_i}\right) \tag{62}$$

$g(\alpha_i)$ is approximated by the function $f(\alpha_i) \propto \alpha_i^{\nu_q}\,\exp(-\beta_q\,\alpha_i)$, by making their respective modes and inflexion points coincide (the computation of the latter involving Newton descent rooting). A proposal distribution is then built as $q(\alpha_i) \propto f(\alpha_i)^N = \mathcal{G}(\alpha_i|N\,\nu_q + 1, N\,\beta_q)$.

A.3 Prior weight of horizontal Markov chains

The expression of the prior weight $\tau_{i,k}$ for the model described in Section 3.3 is given $\forall q = 1,\ldots,l_{\text{frame}}$ as follows:

- $p = 1$

$$\begin{aligned}
\tau_{i,(q,1)} &= \frac{P(\gamma_{i,(q,2)}|\gamma_{i,(q,1)}=1)\;P(\gamma_{i,(q,1)}=1)}{P(\gamma_{i,(q,2)}|\gamma_{i,(q,1)}=0)\;P(\gamma_{i,(q,1)}=0)} \\
&= \begin{cases} \frac{(1-P_{i,11})\,\pi_i}{P_{i,00}\,(1-\pi_i)} = \frac{(1-P_{i,00})}{P_{i,00}} & \text{if } \gamma_{i,(q,2)}=0 \\ \frac{P_{i,11}\,\pi_i}{(1-P_{i,00})\,(1-\pi_i)} = \frac{P_{i,11}}{(1-P_{i,11})} & \text{if } \gamma_{i,(q,2)}=1 \end{cases}
\end{aligned}$$

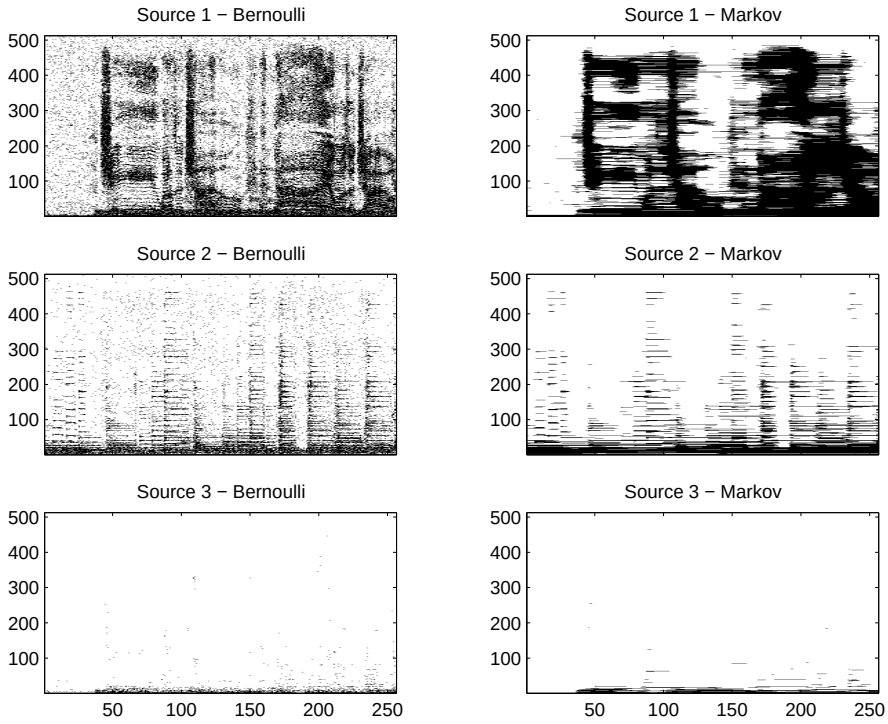

Figure 10: Significance maps of the estimated sources, obtained with Bernoulli priors (left) and horizontal Markov priors (right). Much of the isolated atoms appearing on the left map have been removed in the right map, the latter has been *regularized*.

- $p = 2, \ldots, n_{\text{frame}} - 1$

$$\begin{aligned}
\tau_{i,(q,p)} &= \frac{P(\gamma_{i,(q,p+1)}|\gamma_{i,(q,p)}=1)\ P(\gamma_{i,(q,p)}=1|\gamma_{i,(q,p-1)})}{P(\gamma_{i,(q,p+1)}|\gamma_{i,(q,p)}=0)\ P(\gamma_{i,(q,p)}=0|\gamma_{i,(q,p-1)})} \\
&= \begin{cases}
\frac{(1-P_{i,00})(1-P_{i,11})}{P_{i,00}^2} & \text{if } \gamma_{i,(q,p-1)}=0 \text{ and } \gamma_{i,(q,p+1)}=0 \\
\frac{P_{i,11}}{P_{i,00}} & \text{if } \gamma_{i,(q,p-1)}=0 \text{ and } \gamma_{i,(q,p+1)}=1 \\
\frac{P_{i,11}}{P_{i,00}} & \text{if } \gamma_{i,(q,p-1)}=1 \text{ and } \gamma_{i,(q,p+1)}=0 \\
\frac{P_{i,11}^2}{(1-P_{i,11})(1-P_{i,00})} & \text{if } \gamma_{i,(q,p-1)}=1 \text{ and } \gamma_{i,(q,p+1)}=1
\end{cases}
\end{aligned}$$

- $p = n_{\text{frame}}$

$$\begin{aligned}
\tau_{i,(q,n_{\text{frame}})} &= \frac{P(\gamma_{i,(q,n_{\text{frame}})}=1|\gamma_{i,(q,n_{\text{frame}}-1)})}{P(\gamma_{i,(q,n_{\text{frame}})}=0|\gamma_{i,(q,n_{\text{frame}}-1)})} \\
&= \begin{cases}
\frac{(1-P_{i,00})}{P_{i,00}} & \text{if } \gamma_{i,(q,n_{\text{frame}}-1)}=0 \\
\frac{P_{i,11}}{(1-P_{i,11})} & \text{if } \gamma_{i,(q,n_{\text{frame}}-1)}=1
\end{cases}
\end{aligned}$$

A.4 Update of Markov transition probabilities

We have

$$\begin{aligned} P(\gamma_i|P_{i,00},P_{i,11},\pi_i) &= \prod_{q=1}^{l_{\text{frame}}}\prod_{p=2}^{n_{\text{frame}}} P(\gamma_{i,(q,p)}|\gamma_{i,(q,p-1)},P_{i,00},P_{i,11}) \qquad (63)\\ &\quad \times P(\gamma_{i,(q,1)}|\pi_i) \\ &= P_{i,00}^{\#\gamma_i(00)}\,(1-P_{i,00})^{\#\gamma_i(01)}\;P_{i,11}^{\#\gamma_i(11)} \qquad (64)\\ &\quad \times(1-P_{i,11})^{\#\gamma_i(10)}\;\left(\frac{i-P_{i,00}}{2-P_{i,00}-P_{i,11}}\right)^{\#\gamma_{i,(q,1)}} \\ &\quad \times\left(\frac{i-P_{i,11}}{2-P_{i,00}-P_{i,11}}\right)^{l_{\text{frame}}-\#\gamma_{i,(q,1)}} \end{aligned}$$

where $\#\gamma_i(ij)$ is defined as the cardinality of the set $\{\gamma_{i,(q,p)}=j|\gamma_{i,(q,p-1)}=i, q=1,\ldots,l_{\text{frame}}, p=2,\ldots,n_{\text{frame}}\}$ and $\#\gamma_{i,(q,1)}$ is the cardinality of the set $\{\gamma_{i,(q,1)}=1, q=1,\ldots,l_{\text{frame}}\}$. Hence, we have

$$\begin{aligned} P(P_{i,00}|\gamma_i,P_{i,11},\alpha_{P_{i,00}},\beta_{P_{i,00}}) &\propto P(\gamma_i|P_{i,00},P_{i,11},\pi_i)\;p(P_{i,00}|\alpha_{P_{i,00}},\beta_{P_{i,00}}) \qquad (65)\\ &\propto \frac{\mathcal{B}(P_{i,00}|\#\gamma_i(00)+\alpha_{P_{i,00}},\#\gamma_i(01)+\#\gamma_{i,(q,1)}+\beta_{P_{i,00}})}{(2-P_{i,00}-P_{i,11})^{l_{\text{frame}}}} \end{aligned}$$

$P_{i,00}$ can be updated using a M-H step, for example using the proposal distribution

$$q(P_{i,00}|\gamma_i,\alpha_{P_{i,00}},\beta_{P_{i,00}}) = \mathcal{B}(P_{i,00}|\#\gamma_i(00)+\alpha_{P_{i,00}},\#\gamma_i(01)+\#\gamma_{i,(q,1)}+\beta_{P_{i,00}}) \qquad (66)$$

The acceptance probability $\alpha(P^\star_{i,00}|P_{i,00})$ of candidate $P^\star_{i,00}$ is simply

$$\alpha(P^\star_{i,00}|P_{i,00}) = \left(\frac{2-P_{i,00}-P_{i,11}}{2-P^\star_{i,00}-P_{i,11}}\right)^{l_{\text{frame}}} \qquad (67)$$

However, because of the exponent l_{frame} in (67), the acceptance ratios can stay very low for long periods of time, yielding poorly mixing chains and long burn-in periods. Instead, we found very satisfying in practice to update the transitions probabilities $P_{i,00}$ and $P_{i,11}$ to the modes of their posterior distributions. After calculations of their derivatives, this simply amounts to root polynomials of order two and to choose the root with value lower to one. We favored this latter option in practice. A similar treatment is done for $P_{i,11}$ whose posterior writes

$$\begin{aligned} &p(P_{i,11}|\gamma_i,P_{i,00},\alpha_{P_{i,11}},\beta_{P_{i,11}}) \\ &\qquad\propto \frac{\mathcal{B}(P_{i,11}|\#\gamma_i(11)+\alpha_{P_{i,11}},\#\gamma_i(10)+l_{\text{frame}}-\#\gamma_{i,(q,1)}+\beta_{P_{i,11}})}{(2-P_{i,00}-P_{i,11})^{l_{\text{frame}}}} \qquad (68) \end{aligned}$$

References

[1] K. H. Knuth, "Bayesian source separation and localization," in *SPIE'98: Bayesian Inference for Inverse Problems*, San Diego, Jul. 1998, pp. 147–158.

[2] ——, "A Bayesian approach to source separation," in *Proc. 1st International Workshop on Independent Component Analysis and Signal Separation*, Aussois, France, Jan. 1999, pp. 283–288.

[3] A. Mohammad-Djafari, "A Bayesian approach to source separation," in *Proc. 19th International Workshop on Bayesian Inference and Maximum Entropy Methods (MaxEnt99)*, Boise, USA, Aug. 1999.

[4] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[5] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE. Special issue on blind identification and estimation*, vol. 9, no. 10, pp. 2009–2025, Oct. 1998.

[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. MIT press, 2000, pp. 841–847.

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

[8] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Computation*, 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 Processing Letters*, vol. 4, no. 4, 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 University Press, 2001.

[11] M. Davies and N. Mitianoudis, "A simple mixture model for sparse overcomplete ICA," *IEE Proceedings on Vision, Image and Signal Processing*, Feb. 2004.

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

[13] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Processing*, vol. 51, no. 3, pp. 760–770, Mar. 2003.

[14] S. Mallat, *A wavelet tour of signal processing*. Academic Press, 1998.

[15] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[16] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. R. Statist. Soc. Series B*, vol. B, no. 36, pp. 99–102, 1974.

[17] H. Snoussi and J. Idier, "Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures," *IEEE Trans. Signal Processing*, vol. 54, no. 9, pp. 3257–3269, Sept. 2006.

[18] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, Nov 1984.

[19] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

[20] J. S. Liu, "The collapsed Gibbs sampler with applications to a gene regulation problem," *J. Amer. Statist. Assoc.*, vol. 89, no. 427, pp. 958–966, Sep. 1994.

[21] 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.

[22] C. Févotte and S. Godsill, "A Bayesian approach to blind separation of sparse sources," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2174–2188, Nov. 2006.

[23] J. Geweke, *Variable selection and model comparison in regression*, 5th ed. Oxford Press, 1996, pp. 609–620, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Swith.

[24] P. J. Wolfe, S. J. Godsill, and W.-J. Ng, "Bayesian variable selection and regularisation for time-frequency surface estimation," *J. R. Statist. Soc. Series B*, 2004.

[25] C. Févotte and S. Godsill, "Sparse linear regression in unions of bases via Bayesian variable selection," *IEEE Signal Processing Letters*, vol. 13, no. 7, pp. 441–444, Jul. 2006.

[26] K. Brandenburg, "MP3 and AAC explained," in *Proc. AES 17th Int. Conf. High Quality Audio Coding*, Florence, Italy, Sep 1999.

[27] L. Daudet and M. Sandler, "MDCT analysis of sinusoids: exact results and applications to coding artifacts reduction," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 3, pp. 302–312, May 2004.

[28] M. Davy, S. Godsill, and J. Idier, "Bayesian Analysis of Polyphonic Western Tonal Music," *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2498–2517, Apr. 2006.

[29] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill, "Sparse linear regression with structured priors and application to denoising of musical audio," *IEEE Transactions on Audio, Speech and Language*, in press.

[30] C. Févotte, "Bayesian blind separation of audio mixtures with structured priors," in *Proc. 14th European Signal Processing Conference (EUSIPCO'06)*, Florence, Italy, Sep. 2006.

[31] E. Vincent, R. Gribonval, C. Févotte, and al., "BASS-dB: the blind audio source separation evaluation database," Available on-line, `http://www.irisa.fr/metiss/BASS-dB/`.

[32] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[33] `http://www.tsi.enst.fr/~fevotte/Samples/book_blind_speech_separation/`.

[34] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, no. 11, pp. 1595–1617, 2002, special issue on Image and Video Coding Beyond Standards.

[35] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4133–4145, Nov. 2006.