Variational Bayesian model averaging for audio source separation

Xabier Jaureguiberry, Emmanuel Vincent, Gael Richard

VARIATIONAL BAYESIAN MODEL AVERAGING FOR AUDIO SOURCE SEPARATION

Xabier Jaureguiberry [1*], *Emmanuel Vincent* [2], *Gaël Richard* [1]

[1] Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, 37-39, rue Dareau 75014 Paris, France
[2] Inria, 54600 Villers-lès-Nancy, France

ABSTRACT

Non-negative Matrix Factorization (NMF) has become popular in audio source separation in order to design source-specific models. The number of components of the NMF is known to have a noticeable influence on separation quality. Many methods have thus been proposed to select the best order for a given task. To go further, we propose here to use model averaging. As existing techniques do not allow an effective averaging, we introduce a generative model in which the number of components is a random variable and we propose a modification to conventional variational Bayesian (VB) inference. Experimental results on synthetic data show promising results as our model leads to better separation results and is less computationally demanding than conventional VB model selection.

Index Terms— Variational Bayes, Non-negative Matrix Factorization, Model Averaging, Audio Source Separation

1. INTRODUCTION

Non-negative Matrix Factorization (NMF) has received increasing attention from the research community since its initial proposal [1]. Its ability to recover part-based representation of non-negative data has found numerous applications in different fields such as classification or source separation. In particular, audio source separation has successfully benefited from the development of source models based on NMF principles [2]. Yet the number of components is often assumed to be known whereas it has been shown to have a noticeable influence on the separation results [3]. Several methods based on statistical modelling and model selection principles have been proposed to automatically determine the best number of components according to the data to be processed.

The original formulation of NMF can be seen as a maximum likelihood (ML) problem [4]. ML estimation cannot be used for model selection because the likelihood is always larger for models of higher order and it tends to overfit the data. To overcome these drawbacks, the literature advocates the use of a full Bayesian framework to perform model selection or averaging [5, 6]. In principle, the Bayesian method considers the parameters of a model as random variables with given distributions. This makes it possible to compute the *marginal likelihood* of the model, also known as the *evidence*, by integrating out the likelihood function with respect to the parameters. This marginal likelihood can then be used to select the most likely model or to combine several models. In practice, the computation of the marginal likelihood is often intractable. Approximate inference techniques are required. Amongst the approximating techniques, variational Bayesian (VB) inference has received particular attention as it is computationally efficient [5, 7]. Practically, VB proposes to approximate the true posterior distribution of the parameters by a factored variational distribution in order to compute a lower bound of the marginal likelihood, also named the *free energy*. This free energy is then used to select the most likely model [8, 9].

Lately, VB has been applied to NMF in order to infer the best number of components. These applications to NMF fall into two categories : parametric vs. nonparametric. Parametric methods consist in computing several NMFs with different numbers of components and in comparing their marginal likelihoods. Synthetic tests in [7] have shown that the marginal likelihood is maximum for the number of components which has been used to generate the data. By contrast, nonparametric methods consider a single NMF model but with a potentially infinite number of components. The method in [10] iteratively deactivates the components that are seen as irrelevant. A similar approach has been proposed in [11] without resorting to a VB framework. Both approaches are able to recover the number of components that was used to generate synthetic data.

In order to further improve separation performance, we propose here to average multiple NMFs instead of selecting the one with the most appropriate order. The study in [12] shows that combining several NMFs with fixed fusion weights can improve source separation quality. It also emphasizes that it is worth adapting these weights to the signal to be processed. A straightforward application of model averaging with weights based on the free energy does not lead to an effective averaging as it turns out to select a single model instead of combining several ones. We thus propose in this article a generative model for NMF-based source separation in which the number of components is seen as a random variable. Our framework, which expands the model in [13], takes advantage of parametric methods without tremendously increasing computation time by jointly estimating several NMF models of different orders. Our study also introduces a variational inference framework which differs from conventional VB by implementing a scale factor that controls the entropy of the distribution over the number of components. The model will be first presented in Section 2 and the variational inference framework will be presented in Section 3. Finally, Section 4 will be dedicated to experimental results on synthetic data before concisely concluding in Section 5.

2. GENERATIVE PROBABILISTIC MODEL

Following [13], we aim at modelling the short time Fourier transform (STFT) of an audio signal. Our study is limited to the case of single-channel mixtures composed of two sources but it can be extended to more channels and sources. The mixing equation in frequency bin f and time frame n is thus written as

$$x_{fn} = \mathbf{A}\mathbf{s}_{fn} + \epsilon_{fn} = s_{1,fn} + s_{2,fn} + \epsilon_{fn} \quad (1)$$

where x_{fn} is the mixture STFT coefficient, $s_{1,fn}$ and $s_{2,fn}$ are the two sources that also compose the source vector $\mathbf{s}_{fn} = [s_{1,fn}\, s_{2,fn}]^T$, $\mathbf{A} = [1\, 1]$ is the mixing matrix and ϵ_{fn} represents

*This work was partly supported under the research programme EDiSon3D (ANR-13-CORD-0008-01) funded by ANR, the French State agency for research.

sensor noise. Each source $s_{j,fn}$ is supposed to follow a circularly-symmetric complex normal distribution

$$s_{j,fn} \sim \mathcal{N}\left(0, v_{j,fn}\right) \tag{2}$$

whose variance is expressed by NMF as $v_{j,fn} = \sum_{k=1}^{K_j} w_{j,fk} h_{j,kn}$. $\mathbf{W}_j = \{w_{j,fk}\}_{k=1..K_j}^{f=1..F}$ and $\mathbf{H}_j = \{h_{j,kn}\}_{k=1..K_j}^{n=1..N}$ are the so-called dictionary and activation matrices of the NMF, with F and N being the numbers of frequency bins and of time frames, respectively.

We propose to consider the number of components K_j as a random variable which follows a categorical distribution

$$K_j \sim \text{Cat}\left(\pi_{j1}, ..., \pi_{jm}, ..., \pi_{jM_j}\right) \tag{3}$$

where m indexes the M_j possible number of components $\left\{K_{j1}, ..., K_{jm}, ..., K_{jM_j}\right\}$, each having an *a priori* probability of π_{jm}. The combination of several NMFs of different orders allows us to describe a given source with different resolutions. As a consequence, we assume here that each number of components has its specific NMF parameters so that in the following, they will also be indexed with m. The variance of source j is thus expressed as

$$v_{j,fn} = \sum_{k=1}^{K_{jm}} w_{jm,fk} h_{jm,kn}. \tag{4}$$

Finally, we assume that for each source j and each number of components K_{jm}, the NMF parameters follow a Gamma distribution

$$w_{jm,fk} \sim \Gamma(a, a) \quad , \quad h_{jm,kn} \sim \Gamma(b, b) \tag{5}$$

as in [10] where a and b are hyperparameters to be chosen.

Assuming that sensor noise follows a zero-mean Gaussian distribution of variance σ^2, the likelihood can be formulated as

$$p(\mathbf{X}|\mathbf{S}) = \prod_{n=1}^{N} \prod_{f=1}^{F} \mathcal{N}\left(x_{fn} | \mathbf{A}\mathbf{s}_{fn}, \sigma^2\right) \tag{6}$$

where the notation $\mathbf{X} = \{x_{fn}\}_{f=1..F}^{n=1..N}$ and $\mathbf{S} = \{\mathbf{s}_{fn}\}_{f=1..F}^{n=1..N}$ is used for the sake of readability. Denoting the set of all model parameters as $\mathbf{Z} = \{\mathbf{S}, \mathbf{W}, \mathbf{H}, \mathbf{K}\}$ with $\mathbf{W} = \{\mathbf{W}_{jm}\}_{j=1,2}^{m=1..M_j}$, $\mathbf{H} = \{\mathbf{H}_{jm}\}_{j=1,2}^{m=1..M_j}$ and $\mathbf{K} = \{K_j\}_{j=1,2}$, the joint distribution can be written as

$$p(\mathbf{X}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{S})\, p(\mathbf{S}|\mathbf{W}, \mathbf{H}, \mathbf{K})\, p(\mathbf{W}|\mathbf{K})\, p(\mathbf{H}|\mathbf{K})\, p(\mathbf{K}). \tag{7}$$

3. VARIATIONAL INFERENCE

Estimating the posterior distribution of the model parameters $p(\mathbf{Z}|\mathbf{X})$ leads to intractable calculation. Variational Bayesian inference gives us a way to approximate $p(\mathbf{Z}|\mathbf{X})$ with a factorized variational distribution $q(\mathbf{Z})$. By contrast to [10], we decided to keep the conditioning of $\mathbf{W}$ and $\mathbf{H}$ on $\mathbf{K}$ so that the distribution of interest $q(\mathbf{S})$ is the only distribution to be approximated. The factorized variational distribution is thus expressed as

$$q(\mathbf{Z}) = \prod_{i=1}^{I} q_i(\mathbf{Z}_i) = q(\mathbf{S})\, q(\mathbf{W}|\mathbf{K})\, q(\mathbf{H}|\mathbf{K})\, q(\mathbf{K}). \tag{8}$$

3.1. Maximizing the free energy

In VB inference, approximating $p(\mathbf{Z}|\mathbf{X})$ to $q(\mathbf{Z})$ is equivalent to maximizing the free energy [13]. The strategy consists in iteratively maximizing the free energy with respect to each distribution q_i in (8). The particular form of our variational distribution results in the following expressions of the distributions which maximize the free energy:

$$\log q^*(\mathbf{S}) = \mathbb{E}_{\mathbf{Z}\setminus\mathbf{S}}\left[\log p(\mathbf{X}, \mathbf{Z})\right] + \text{const} \tag{9}$$

$$\log q^*(\mathbf{W}|\mathbf{K}) = \mathbb{E}_{\mathbf{Z}\setminus\{\mathbf{W},\mathbf{K}\}}\left[\log p(\mathbf{X}, \mathbf{Z})\right] + \text{const} \tag{10}$$

$$\log q^*(\mathbf{H}|\mathbf{K}) = \mathbb{E}_{\mathbf{Z}\setminus\{\mathbf{H},\mathbf{K}\}}\left[\log p(\mathbf{X}, \mathbf{Z})\right] + \text{const} \tag{11}$$

$$\begin{aligned}\log q^*(\mathbf{K}) = &\mathbb{E}_{\mathbf{Z}\setminus\mathbf{K}}\left[\log p(\mathbf{X}, \mathbf{Z})\right] - \mathbb{E}_{\mathbf{W}}\left[\log q(\mathbf{W}|\mathbf{K})\right] \\ &- \mathbb{E}_{\mathbf{H}}\left[\log q(\mathbf{H}|\mathbf{K})\right] + \text{const}\end{aligned} \tag{12}$$

where $\mathbb{E}_{\mathbf{Z}\setminus\mathbf{Z}_i}\left[\log p(\mathbf{X}, \mathbf{Z})\right]$ denotes the expectation of the joint distribution (7) over all model parameters $\mathbf{Z}$ except $\mathbf{Z}_i$.

3.2. Lower bounding the free energy

In order to find the update rules of the model parameters, we need to compute the expectations in (9) – (12). Amongst these terms, $\mathbb{E}_{\mathbf{Z}_*}\left[\log p(\mathbf{S}|\mathbf{W}, \mathbf{H}, \mathbf{K})\right]$, in which $\mathbf{Z}_*$ denotes any subset of $\mathbf{Z}$, is intractable. At first, it is worth noting that

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}_*}\left[\log p(\mathbf{S}|\mathbf{W}, \mathbf{H}, \mathbf{K})\right] &= \sum_{j,fn} \mathbb{E}_{\mathbf{Z}_*}\left[\log p(s_{j,fn}|\mathbf{W}_j, \mathbf{H}_j, K_j)\right] \\ &= \sum_{jm,fn} q(K_{jm})\, \mathbb{E}_{\mathbf{Z}_*\setminus K_j}\left[\log p(s_{j,fn}|\mathbf{W}_{jm}, \mathbf{H}_{jm})\right]\end{aligned}$$

Following [10], we lower bound $\mathbb{E}_{\mathbf{Z}_*\setminus K_j}\left[\log p(s_{j,fn}|\mathbf{W}_{jm}, \mathbf{H}_{jm})\right]$ as follows:

$$\begin{aligned}&\mathbb{E}_{\mathbf{Z}_*\setminus K_j}\left[\log p(s_{j,fn}|\mathbf{W}_{jm}, \mathbf{H}_{jm})\right] \geq -\log\pi \\ &\quad - \log\omega_{jm,fn} + 1 - \frac{1}{\omega_{jm,fn}} \sum_{k=1}^{K_{jm}} \mathbb{E}_{\mathbf{Z}_*\setminus K_j}\left[w_{jm,fk} h_{jm,kn}\right] \\ &- \mathbb{E}_{\mathbf{Z}_*\setminus K_j}\left[|s_{j,fn}|^2\right] \sum_{k=1}^{K_{jm}} \phi^2_{jm,fn,k} \mathbb{E}_{\mathbf{Z}_*\setminus K_j}\left[\frac{1}{w_{jm,fk} h_{jm,kn}}\right].\end{aligned}$$

for any $\omega_{jm,fn} \geq 0$ and $\phi_{jm,fn,k} \geq 0$ s.t. $\sum_{k=1}^{K_{jm}} \phi_{jm,fn,k} = 1$.

This lower bound is tightened by zeroing its derivative w.r.t. $\omega_{jm,fn}$ and $\phi_{jm,fn,k}$ which leads to the following expressions :

$$\omega_{jm,fn} = \sum_{k=1}^{K_j m} \mathbb{E}_{\mathbf{Z}_*\setminus K_j}\left[w_{jm,fk} h_{jm,kn}\right] \tag{13}$$

$$\phi_{jm,fn,k} = \frac{1}{C_{jm,fn}} \mathbb{E}_{\mathbf{Z}_*\setminus K_j}\left[\frac{1}{w_{jm,fk} h_{jm,kn}}\right]^{-1} \tag{14}$$

with $C_{jm,fn} = \sum_{k=1}^{K_{jm}} \mathbb{E}_{\mathbf{Z}_*\setminus K_j}\left[\frac{1}{w_{jm,fk} h_{jm,kn}}\right]^{-1}$.

3.3. Variational updates

When needed, the expectations of $\log p(\mathbf{S}|\mathbf{W}, \mathbf{H}, \mathbf{K})$ in (9) – (12) are replaced by their parametric lower bounds. The variational distribution of the sources is identified to a bivariate Gaussian distribution $q(\mathbf{s}_{fn}) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{s},fn}, \boldsymbol{\Sigma}_{\mathbf{s},fn}\right)$ with parameters

$$\boldsymbol{\mu}_{\mathbf{s},fn} = \boldsymbol{\Sigma}_{\mathbf{s},fn} \mathbf{A} \frac{1}{\sigma^2} x_{fn} \ , \ \boldsymbol{\Sigma}_{\mathbf{s},fn} = \left(\mathbf{C}_{fn}^{-1} + \frac{1}{\sigma^2}\mathbf{J}\right)^{-1} \tag{15}$$

where $\mathbf{J}$ is a matrix of ones of size 2×2, $\mathbf{C}_{fn}^{-1} = \text{diag}(C_{j,fn}^{-1})_{j=1,2}$ and

$$C_{j,fn}^{-1} = \sum_{m=1}^{M_j} q(K_{jm}) C_{jm,fn}^{-1}. \tag{16}$$

The variational distributions of the NMF parameters are identified to generalized inverse Gaussian (GIG) distributions [10] which are controlled by three parameters τ, ρ and γ. The updates of these parameters for the matrix $\mathbf{W}_{jm} = \{w_{jm,fk}\}_{f=1..F}^{k=1..K_{jm}}$ of source j for the number of components K_{jm} are given by:

$$\begin{aligned}\boldsymbol{\tau}^{\mathbf{W}}_{jm} &= \mathbb{E}\left[\frac{1}{\mathbf{W}_{jm}}\right]^{.2} \circ \left[\left(\mathbb{E}\left[|\mathbf{S}_j|^{.2}\right] \circ \mathbf{C}_{jm}^{.-2}\right)\left(\mathbb{E}\left[\frac{1}{\mathbf{H}_{jm}}\right]^{.-1}\right)^T\right] \\ \boldsymbol{\rho}^{\mathbf{W}}_{jm} &= a + \mathbb{E}\left[\mathbf{V}_{jm}\right]^{.-1} \mathbb{E}\left[\mathbf{H}_{jm}\right]^T \quad , \quad \boldsymbol{\gamma}^{\mathbf{W}}_{jm} = a\end{aligned} \tag{17}$$

where the notation $\circ$ denotes the Hadamard product, $\mathbf{M}^{.x}$ denotes element-wise exponentiation and $\mathbf{M}^T$ denotes the transpose of matrix $\mathbf{M}$. $\mathbf{V}_{jm}$ is the variance of source j for the number of components K_{jm}, that is to say the product $\mathbf{W}_{jm}\mathbf{H}_{jm}$. $\mathbf{C}_{jm}$ is the matrix composed of the coefficients $C_{jm,fn}$ defined in (14). As part of the

exponential family, the GIG distribution over $\mathbf{W}_{jm}$ is equivalently determined by its parameters $\boldsymbol{\tau}^{\mathbf{W}}_{jm}$, $\boldsymbol{\rho}^{\mathbf{W}}_{jm}$ and $\boldsymbol{\gamma}^{\mathbf{W}}_{jm}$ or by its statistics $\mathbb{E}[\mathbf{W}_{jm}]$, $\mathbb{E}\left[\frac{1}{\mathbf{W}_{jm}}\right]$ and $\mathbb{E}[\log \mathbf{W}_{jm}]$. The inference scheme thus alternates between estimating the statistics and the parameters of the distribution as in an expectation-maximization algorithm. Note that the same update rules can be found for $\mathbf{H}_{jm}$ by replacing and reordering the terms accordingly.

Finally, the log-posterior distribution $\log \tilde{q}(K_{jm})$ of the number of components K_j is computed as the sum of the terms given in (12). Besides, it is worth noting that when considering a single number of components ($M_j = 1$), our model is equivalent to the model in [13].

3.4. Posterior over the model order

To obtain the posterior distribution of K_j, it is necessary to take the exponential of $\log \tilde{q}(K_{jm})$ and to normalize it so that $\sum_{m=1}^{M_j} q(K_{jm}) = 1$. Preliminary tests have shown that the sum of all terms in (12) gives large values, which results in one $q(K_{jm})$ being equal to one and the others being equal to zero when taking the exponential. This means that conventional VB inference leads to model selection rather than model averaging. In order to avoid it, we propose to scale the log-posterior by a factor β before computing the exponential. This is equivalent to penalizing the entropy of the distribution $q(K_j)$ in a way similar to [14]. The posterior probability is thus computed as

$$q(K_{jm}) \propto \exp\left(\frac{\log \tilde{q}(K_{jm})}{\beta}\right). \tag{18}$$

Thus, small values of β will favour peaky distributions with one $q(K_{jm})$ close to 1, whereas higher values of β will result in a more uniform distribution over K_j.

4. EXPERIMENTS

We propose to evaluate our model on an audio source separation task. To do so, we rely on the PASCAL CHiME corpus [15] which features recordings of real domestic noise and speech utterances from diverse speakers.

4.1. Synthetic data generation

We randomly selected one speaker of the database and ten seconds of background noise. We learned an NMF speaker model by concatenating 250 utterances of the selected speaker and estimating $\mathbf{W}_{1m}$ and $\mathbf{H}_{1m}$ following a standard maximum-likelihood scheme. We chose seven different numbers of components so that $K_{1m} = 2^m$ with $m = 1..7$. We retained the seven dictionaries $\{\mathbf{W}_{1m}\}_{m=1..7}$ as seven models of the same speaker, hence describing the same spectral content but at different levels of details. For instance, $\mathbf{W}_{11}$ will give a rough description of the spectral characteristics of the speaker whereas $\mathbf{W}_{17}$ will be a much more detailed description but with potential redundancies. We proceeded the same way to learn an NMF background model. This time, we retained a single model with 16 components. For each learning, we used a STFT with half-overlapping windows of 2048 samples each.

Once the models for both speech and background were learned, we generated for each model order K_{1m} several mixtures $\mathbf{X}$, with 300 time frames each, according to our generative model. The dictionary $\mathbf{W}_{1m}$ of the speaker source and the dictionary $\mathbf{W}_2$ of the background source were fixed to their learned values. The corresponding activation matrices $\mathbf{H}_{1m}$ and $\mathbf{H}_2$ were randomly generated according to the Gamma distribution in (5) with $b = 0.2$. Both the speaker and background sources $\mathbf{S}_1$ and $\mathbf{S}_2$ were randomly generated according to the distribution defined in (2). Finally, the mixing equation (1) was used to generate the observation $\mathbf{X}$ at different signal-to-noise ratios (SNRs). Here, we chose six different values of SNR from -6 dB to 9 dB by step of 3 dB. The sensor noise ϵ_{fn} in (1) has been chosen of constant variance $\sigma^2 = 10^{-6}$. As a consequence, we generated a total of 42 synthetic observations $\mathbf{X}$, *i.e.*, for six different SNRs and seven model orders.

4.2. Estimation and separation

For each generated example, we propose to compare *VB selection* and *VB fusion*. VB selection consists in computing seven single-order NMFs with $K_{1m} = 2^m$ and in retaining the one which gives the highest bound to the free energy. Firstly, for a given number of components K_{1m}, a single-order speaker model is learned on 250 utterances of the same speaker as in the generation step yet the utterances differ from those used to generate the data. This choice is motivated by [15] in which the training data are distinct from the development and test data in order to simulate real domotic scenarios. We then resort to the model introduced in Section 2 with $M_1 = M_2 = 1$ and to the variational inference scheme exposed in Section 3. The distribution over $\mathbf{W}_{1m}$ is set to be a Dirac according to the single-order NMF model previously learned. The distribution over $\mathbf{W}_2$ is initialized by its statistics $\mathbb{E}[\mathbf{W}_2]$ and $\mathbb{E}[1/\mathbf{W}_2]$, the statistic $\mathbb{E}[\log \mathbf{W}_2]$ being unused in the variational updates. The terms of $\mathbb{E}[\mathbf{W}_2]$ are drawn according to the Gamma prior in (5) with $a = 0.2$ and $\mathbb{E}[1/\mathbf{W}_2]$ is set to be equal to $\mathbb{E}[\mathbf{W}_2]^{\cdot -1}$. In the same manner, $\mathbb{E}[\mathbf{H}_{1m}]$, $\mathbb{E}[1/\mathbf{H}_{1m}]$, $\mathbb{E}[\mathbf{H}_2]$ and $\mathbb{E}[1/\mathbf{H}_2]$ are initialized as constants. The distribution $q(\mathbf{W}_{1m})$ is fixed whereas $q(\mathbf{W}_2)$, $q(\mathbf{H}_{1m})$ and $q(\mathbf{H}_2)$ are to be estimated. We used 50 iterations and at the end, the estimated sources are computed as the expectation $\boldsymbol{\mu}_{\mathbf{s},fn}$ of the posterior distribution in (15). As we only consider single-order NMF models, the update of $q(\mathbf{K})$ in (12) is of no use here for both the speaker and background models.

On the contrary, *VB fusion* uses the full model we exposed in Section 2, including the update of $q(\mathbf{K})$. Instead of considering a single speaker model, the framework jointly takes into account the seven learned speaker models. The term fusion here refers to the fact that the variance of the speaker source is now the combination of several NMFs of different orders because of model averaging. The prior probability of the number of components of the speaker source is set to be uniform, *i.e.*, $p(K_{1m}) = 1/7$ with $K_{1m} = 2^m$ and $m = 1..7$. The corresponding distributions over $\mathbf{W}_{1m}$ and $\mathbf{H}_{1m}$ and the distributions related to the background source are initialized as in the selection case above. Note that preliminary tests have shown that with $\beta = 1$, VB fusion leads to similar results as VB selection. Indeed, the expression in (12) thus results in one number of components having a posterior probability of one and the others having zero probability. Therefore, in order to compute the posterior probability of the number of components, we weight its entropy by $\beta = 10^4$, a value which has been found to work well in practice.

4.3. Results

The separation quality is evaluated by the signal-to-distortion ratio (SDR) expressed in decibels [16]. Fig. 1 shows the SDR of the speaker source for VB selection and VB fusion. The best single NMF separation and the mean of the seven separated signals are also evaluated for indicative purposes. They are denoted as *oracle selection* and *fusion by mean* in reference to [12]. Note that oracle selection is not reachable in practice as it requires the knowledge of the original true speaker source. The SDRs are averaged over the seven numbers of components used at generation for each SNR. The corresponding values as well as the average computation time are reported in Table 1.

The comparison of VB selection and oracle selection shows that VB selection fails to retain the number of components which gives

	SNR							Average
	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Average	time (ms)
VB selection	3.77	4.36	5.20	5.72	6.19	6.40	5.27	172.7
VB fusion with $\beta = 10^4$	4.24	4.90	5.47	6.14	6.76	7.21	5.79	60.8
Oracle selection	4.35	5.04	5.69	6.45	7.14	7.77	6.07	172.7
Fusion by mean	3.81	4.57	5.22	5.91	6.53	7.04	5.52	172.7

Table 1. Average SDR (dB) for VB selection, VB fusion, oracle selection and fusion by mean, for each SNR and average computation time

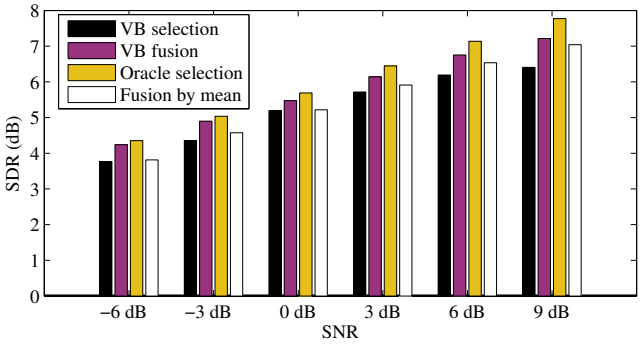

Fig. 1. Average SDR for VB selection, VB fusion, oracle selection and fusion by mean, for each SNR

the best SDR in our study case. On average, VB selection underperforms oracle selection by 0.8 dB. However, the proposed VB fusion scheme gives significantly better results than VB selection as it outperforms the latter by 0.5 dB on average. Moreover, VB fusion outperforms the simple fusion by mean. Finally, besides being more efficient in terms of SDR, VB fusion is also less time consuming than VB selection as it is 2.8 times faster. Indeed, VB fusion uses less parameters than VB selection. In particular, there is an unique background model in VB fusion whereas VB selection requires the estimation of a background model per number of components.

5. CONCLUSION

We introduced a generative model dedicated to NMF-based source separation. Using VB inference together with a parametric lower bound of the marginal likelihood, our model allows us to describe a source as the combination of several NMFs of different orders, in a way similar to model averaging. Our experimental results on synthetic data show that our model gives better separation results than VB model selection and almost reaches oracle selection results. Our VB fusion approach also turns out to be more computationally efficient than VB selection. Future work will focus on the study of the influence and determination of β as well as on the application of our generative model to real data.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[2] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[3] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2007, vol. 1, pp. I–65–68.

[4] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2008, pp. 1825–1828.

[5] H. Attias, "A variational Bayesian framework for graphical models," *Advances in neural information processing systems*, vol. 12, no. 1-2, pp. 209–215, 2000.

[6] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical science*, pp. 382–401, 1999.

[7] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, 2009, Article ID 785152.

[8] A. Corduneanu and C. M. Bishop, "Variational Bayesian model selection for mixture distributions," in *Proc. of the 8th International Workshop on Artificial Intelligence and Statistics*, 2001, pp. 27–34.

[9] J. M. Bernardo et al., "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Proc. of Valencia International Meeting on Bayesian Statistics*, 2002, pp. 453–462.

[10] M. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. of International Conference on Machine Learning (ICML)*, 2010, pp. 439–446.

[11] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Proc. of Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2009.

[12] X. Jaureguiberry, G. Richard, P. Leveau, R. Hennequin, and E. Vincent, "Introducing a simple fusion framework for audio source separation," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[13] K. Adilŏglu and E. Vincent, "Variational Bayesian inference for source separation and robust feature extraction," Tech. Rep. RT-0428, Inria, 2012.

[14] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, no. 5, pp. 1155–1182, 1999.

[15] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: datasets, tasks and baselines," in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2013, pp. 126–130.

[16] 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, 2006.