A VARIATIONAL BAYES APPROACH TO THE UNDERDETERMINED BLIND SOURCE SEPARATION WITH AUTOMATIC DETERMINATION OF THE NUMBER OF SOURCES

Jalil Taghia, Nasser Mohammadiha, Arne Leijon

Sound and Image Processing Lab., KTH Royal Institute of Technology, Sweden,
{taghia, nmoh, leijon}@kth.se

ABSTRACT

In this paper, we propose a variational Bayes approach to the underdetermined blind source separation and show how a variational treatment can open up the possibility of determining the actual number of sources. The procedure is performed in a frequency bin-wise manner. In every frequency bin, we model the time-frequency mixture by a variational mixture of Gaussians with a circular-symmetric complex-Gaussian density function. In the Bayesian inference, we consider appropriate conjugate prior distributions for modeling the parameters of this distribution. The learning task consists of estimating the hyper-parameters characterizing the parameter distributions for the optimization of the variational posterior distribution. The proposed approach requires no prior knowledge on the number of sources in a mixture.

Index Terms— blind source separation, variational Bayesian approach, number of sources, variational mixture of Gaussians

1. INTRODUCTION

Most blind source separation (BSS) approaches rely on the assumption that the number of mixed sources in an observation mixture is known. This assumption restricts the application of BSS in real-world scenarios. In this paper, we propose a variational Bayes approach that requires no prior knowledge on the number of sources. In [1], motivated by the line orientation idea [2], the observation mixture is modeled by a mixture of circular-symmetric complex-Gaussian distributions where expectation maximization (EM) algorithm is employed to estimate parameters of the density function and posterior probabilities. This approach needs to know the number of components in advance. On the other hand, the EM algorithm is shown to be sensitive to the initialization [3] because, depending on starting values, it may converge to a local maximum of the observed-data likelihood-function and provide only a suboptimal solution. [1, 4, 5] can be regarded as the pioneer work for solving the underdetermined BSS with EM algorithm.

The model uncertainty can be taken into account in a Bayesian framework by considering distribution of parameters instead of parameter point estimates [6, 7]. Bayesian approaches do not suffer from overfitting which can be considered as one of the substantial advantages of Bayesian methods over maximum-likelihood ones. In this paper, similar to [1], we model the observation mixture by a variational mixture of circular-symmetric complex-Gaussian distributions. We define proper conjugate prior distributions for modeling the parameters of the mixture model and use variational Bayes treatment [7] for estimating hyper-parameters characterizing the parameter distributions. In the experimental result, we compare the separation performance of the proposed variational Bayes approach with the maximum-likelihood based EM realization of that [1] which will be called the EM-based approach, afterwards. We show that the proposed approach has competitive separation performance while it requires no prior knowledge on the number of sources and with little overhead in terms of the computational complexity.

This work was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699.

2. PROBLEM FORMULATION

Let $s_1(t), \ldots, s_Q(t)$ be desired sources and $y_1(t), \ldots, y_D(t)$ be observation mixtures, where Q and D indicate the number of sources and the number of observations, respectively. Assuming a convolutive mixture model, the observation $y_d(t)$ is given by $y_d(t) = \sum_{q=1}^{Q} \sum_l h_{dq}(l) s_q(t-l)$, where $h_{dq}(l)$ represents the impulse response from source q to microphone d. The convolutive mixture model $y_d(t)$ is transformed to the time-frequency representation by using the short time Fourier transform and can be approximated as an instantaneous mixture model at each frequency bin as $y_d(n,f) = \sum_{q=1}^{Q} h_{dq}(f) s_q(n,f)$, where $s_q(n,f)$ is the time-frequency representation of $s_q(t)$, f indicates the frequency bin, and n indicates the time frame. In vector notation, it can be written as $\mathbf{y}(n,f) = \sum_{q=1}^{Q} \mathbf{h}_q(f) s_q(n,f)$ which under sparsity assumption can be expressed by $\mathbf{y}(n,f) = \mathbf{h}_{q'}(f) s_{q'}(n,f)$. Subscript q' is the index of the dominant source.

To remove the source amplitude effect, the observations $\mathbf{y}(n,f)$ are normalized such that they have a unit norm. It can be achieved by $\mathbf{x}(n,f) = \frac{\mathbf{y}(n,f)}{\|\mathbf{y}(n,f)\|}$. A pre-whitening, [8, 1], is performed by multiplying $\mathbf{x}(n,f)$ by the whitening matrix $\mathbf{W}$, as: $\mathbf{x}(n,f) \leftarrow \mathbf{W}\mathbf{x}(n,f)$, where $\mathbf{W} = \sqrt{\mathbf{A}}\mathbf{G}^H$. $\mathbf{G}$ and $\mathbf{A}$ are calculated from eigenvalue decomposition of the correlation matrix $\mathrm{E}[\mathbf{x}\mathbf{x}^H] = \mathbf{G}\mathbf{A}\mathbf{G}^H$. The normalization procedure is performed one more time after whitening. In the rest of the paper, we omit the frequency-bin index since all procedure is performed in a frequency bin-wise manner.

3. MODEL DESCRIPTION

In the mixture model, for each observation $\mathbf{x}_n$, there is a corresponding latent variable (indication vector) $\mathbf{z}_n$ which forms a 1-of-K binary vector with elements z_{nk}. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ denote the observation set in a particular frequency bin, and $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ denote the latent variables, where N indicates the whole number of time frames. The conditional distribution of $\mathbf{Z}$ given the mixing coefficients $\boldsymbol{\gamma} = \{\gamma_k\}$ can be expressed by

$$p(\mathbf{Z}|\boldsymbol{\gamma}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \gamma_k^{z_{nk}}. \tag{1}$$

We express the conditional distribution of $\mathbf{X}$ given the latent variables $\mathbf{Z}$ and component parameters $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\lambda} = \{\lambda_k\}$ by

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \widetilde{\mathbb{N}}_c(\mathbf{x}_n|\boldsymbol{\mu}_k, \lambda_k^{-1})^{z_{nk}}, \tag{2}$$

where, motivated by [1, 2], $\widetilde{\mathbb{N}}_c(\mathbf{x}_n|\boldsymbol{\mu}_k, \lambda_k^{-1})$ denotes a circular-symmetric complex-Gaussian density function

$$\widetilde{\mathbb{N}}_c(\mathbf{x}_n|\boldsymbol{\mu}_k, \lambda_k^{-1}) = \frac{1}{(\pi\lambda_k^{-1})^{D-1}} e^{-\lambda_k \|\mathbf{x}_n - (\boldsymbol{\mu}_k^H \mathbf{x}_n)\boldsymbol{\mu}_k\|^2}, \tag{3}$$

where $\boldsymbol{\mu}_k$ is the centroid with unit norm, $\boldsymbol{\mu}_k^H \boldsymbol{\mu}_k = 1$, and λ_k is the precision which is scalar and the same for all k. $(\boldsymbol{\mu}_k^H \mathbf{x}_n)\boldsymbol{\mu}_k$ is

the orthogonal projection of $\mathbf{x_n}$ onto the subspace spanned by $\boldsymbol{\mu}_k$, hence, the distance $\|\mathbf{x}_n - (\boldsymbol{\mu}_k^H \mathbf{x}_n)\boldsymbol{\mu}_k\|^2$ determines the dependency of $\mathbf{x}_n$ to the k^{th} class. In the next step, we consider priors over the parameters $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\gamma}$. A Dirichlet prior distribution is introduced over the mixing coefficients $\boldsymbol{\gamma}$ as

$$p(\boldsymbol{\gamma}) = \text{Dir}(\boldsymbol{\gamma}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \gamma_k^{\alpha_0 - 1}, \tag{4}$$

where α_0 is chosen the same for all components and $C(\boldsymbol{\alpha}) = \frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha_1 \ldots \alpha_K)}$, $\hat{\alpha} = \sum_{k=1}^{K} \alpha_k$. For small values of α_0, the posterior distribution will be influenced mostly by the data and not by the prior. A Gaussian-Gamma prior is introduced to govern the mean and precision of each Gaussian component, given by

$$p(\boldsymbol{\mu}, \boldsymbol{\lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) = \prod_{k=1}^{K} \mathbb{N}_c\big(\boldsymbol{\mu}_k|\mathbf{m_0}, (\lambda_k\beta_0\mathbf{I})^{-1}\big)\mathbb{G}(\lambda_k|a_0, b_0), \tag{5}$$

where $\mathbb{N}_c(\boldsymbol{\mu}_k|\mathbf{m}_0, (\lambda_k\beta_0\mathbf{I})^{-1})$ is a circular-symmetric complex-Gaussian density fuction with mean value $\mathbf{m}_0$ and the precision $(\lambda_k\beta_0\mathbf{I})^{-1}$ (β_0 is a scalar and $\mathbf{I}$ is the identity matrix) of the form

$$\mathbb{N}_c\big(\boldsymbol{\mu}_k|\mathbf{m_0}, (\lambda_k\beta_0\mathbf{I})^{-1}\big) = \frac{1}{\big(\pi(\lambda_k\beta_0)^{-1}\big)^D} e^{-\lambda_k(\boldsymbol{\mu}_k - \mathbf{m_0})^H \beta_0 \mathbf{I}(\boldsymbol{\mu}_k - \mathbf{m_0})}. \tag{6}$$

It is worth noting that considering such a conjugate distribution for modeling $p(\boldsymbol{\mu}, \boldsymbol{\lambda})$ may not be the best possible choice since we require $\boldsymbol{\mu}_k^H \boldsymbol{\mu}_k = 1$ in (3). However, we show that it can be a good approximation to be used as a conjugate prior. $\mathbb{G}(\lambda_k|a_0, b_0)$ is a Gamma density function with the shape parameter a_0 and the scale parametr b_0, given by

$$\mathbb{G}(\lambda_k|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda_k^{a_0 - 1} e^{-b_0\lambda_k}, \tag{7}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

4. OPTIMIZATION OF THE VARIATIONAL POSTERIOR DISTRIBUTION

We employ the variational Bayes approach [5]. Having only the observation set $\mathbf{X}$ observed, the joint distribution of the observation set, latent variables, and component parameters is given by

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\lambda})p(\mathbf{Z}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\boldsymbol{\mu}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}). \tag{8}$$

The variational posterior distribution which factorizes the latent variables and the component parameters can be written by

$$q(\mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = q(\mathbf{Z})q(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda}). \tag{9}$$

The optimization of the variational posterior distribution $q(\mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ consists of optimization of the variational posterior distribution of the latent variables $q(\mathbf{Z})$ and component parameters $q(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda})$. Starting with the optimization of $q(\mathbf{Z})$, the log of the optimized factor [1] $q^*(\mathbf{Z})$ is given by

$$\begin{aligned} \ln q^*(\mathbf{Z}) &= \text{E}_{\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\lambda}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda})] + \text{const} \\ &= \text{E}_{\boldsymbol{\gamma}}[\ln p(\mathbf{Z}|\boldsymbol{\gamma})] + \text{E}_{\boldsymbol{\mu},\boldsymbol{\lambda}}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\lambda})] + \text{const}, \end{aligned} \tag{10}$$

[1] Let $p(\mathbf{Y}, \boldsymbol{\theta})$ denote the joint distribution of the probabilistic model of the set of observed variables $\mathbf{Y}$ and the set of all latent variables and parameters $\boldsymbol{\theta}$ where $q(\boldsymbol{\theta}) = \prod_{i=1}^{I} q_i(\boldsymbol{\theta})$. The general expression for the optimal solution $q_j^*(\boldsymbol{\theta}_j)$ is given by $\ln q_j^*(\boldsymbol{\theta}_j) = \text{E}_{i \neq j}[\ln p(\mathbf{Y}, \boldsymbol{\theta})]$ [7].

where the superscript (*) is used to show the optimized factor. By substituting (1) and (2) in (10) and including terms which are independent of $\mathbf{Z}$ in the constant term, we obtain

$$\ln q^*(\mathbf{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const}, \tag{11}$$

where

$$\begin{aligned} \ln \rho_{nk} = &-(D-1)\ln \pi + (D-1)\text{E}_{\lambda_k}[\ln \lambda_k] + \\ &\text{E}_{\boldsymbol{\gamma}_k}[\ln \boldsymbol{\gamma}_k] - \text{E}_{\boldsymbol{\mu}_k,\lambda_k}[\lambda_k \|\mathbf{x}_n - (\boldsymbol{\mu}_k^H \mathbf{x}_n)\boldsymbol{\mu}_k\|^2]. \end{aligned} \tag{12}$$

Let us rewrite (11) as $q^*(\mathbf{Z}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$ and normalize it so that for each value n, the quantities z_{nk} are binary and sum to one over all values of k. Hence, we get

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \xi_{nk}^{z_{nk}}, \tag{13}$$

which is a multinomial distribution with

$$\xi_{nk} = \frac{\rho_{nk}}{\sum_{k=1}^{K} \rho_{nk}}. \tag{14}$$

ξ_{nk} are called responsibilities, and it is worth noting that these responsibilities are real and sum to unit, thus, for the multinomial distribution $q^*(\mathbf{Z})$, we have $\text{E}_{\mathbf{Z}}[z_{nk}] = \xi_{nk}$.

Next, we consider the optimization of the variational posterior distribution of the component parameters $q(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ in (9). The log of the optimized factor $q^*(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ is given by

$$\begin{aligned} \ln q^*(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \text{E}_{\mathbf{Z}}[&\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \\ &+ \ln p(\mathbf{Z}|\boldsymbol{\gamma}) + \ln p(\boldsymbol{\gamma}) + \ln p(\boldsymbol{\mu}, \boldsymbol{\lambda})]. \end{aligned} \tag{15}$$

Considering (2), (1), (4), and (5) in (15), we obtain

$$\begin{aligned} \ln q^*(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = &\sum_{n=1}^{N} \sum_{k=1}^{K} \xi_{nk} \ln \widetilde{\mathbb{N}}_c(\mathbf{x}_n|\boldsymbol{\mu}_k, \lambda_k^{-1}) + \\ &\sum_{n=1}^{N} \sum_{k=1}^{K} \xi_{nk} \ln \gamma_k + \ln \text{Dir}(\boldsymbol{\gamma}|\boldsymbol{\alpha}_0) + \\ &\sum_{k=1}^{K} \ln \mathbb{N}_c\big(\boldsymbol{\mu}_k|\mathbf{m_0}, (\lambda_k\beta_0\mathbf{I})^{-1}\big) + \ln \mathbb{G}(\lambda_k|a_0, b_0). \end{aligned} \tag{16}$$

The right hand side of (16) includes terms which involve either $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ or $\boldsymbol{\gamma}$, hence, $q(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ can be factorized as $q(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = q(\boldsymbol{\gamma})q(\boldsymbol{\mu}, \boldsymbol{\lambda})$. First, we start with optimization of the variational posterior distribution of the mixture weights $q(\boldsymbol{\gamma})$. By identifying those terms in (16) that only involve $\boldsymbol{\gamma}$, we get

$$\ln q^*(\boldsymbol{\gamma}) = \sum_{k=1}^{K} \big((\alpha_0 - 1) + \sum_{n=1}^{N} \xi_{nk}\big) \ln \boldsymbol{\gamma}_k + \text{const}. \tag{17}$$

Thus, $q^*(\boldsymbol{\gamma})$ is given by

$$q^*(\boldsymbol{\gamma}) = \text{Dir}(\boldsymbol{\gamma}|\boldsymbol{\alpha}), \tag{18}$$

where

$$\alpha_k = \alpha_0 + \sum_{n=1}^{N} \xi_{nk}. \tag{19}$$

Next, we consider the optimization of the second term $q(\boldsymbol{\mu}, \boldsymbol{\lambda})$ in the variational posterior distribution $q(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\lambda})$. $q(\boldsymbol{\mu}, \boldsymbol{\lambda})$ can be factorized as $q(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \lambda_k)$. In order to derive $q^*(\boldsymbol{\mu}_k, \lambda_k)$, we identify the terms in (16) which depend only on $\boldsymbol{\mu}_k$ and λ_k, hence

$$\begin{aligned} \ln q^*(\boldsymbol{\mu}_k, \lambda_k) = &\sum_{n=1}^{N} \xi_{nk} \ln \widetilde{\mathbb{N}}_c(\mathbf{x}_n|\boldsymbol{\mu}_k, \lambda_k^{-1}) + \\ &\ln \mathbb{N}_c\big(\boldsymbol{\mu}_k|\mathbf{m_0}, (\lambda_k\beta_0\mathbf{I})^{-1}\big) + \ln \mathbb{G}(\lambda_k|a_0, b_0). \end{aligned} \tag{20}$$

The variational posterior distribution $q^*(\boldsymbol{\mu}_k, \lambda_k)$, by using the Bayes rule, can be factorized as

$$q^*(\boldsymbol{\mu}_k, \lambda_k) = q^*(\boldsymbol{\mu}_k|\lambda_k)q^*(\lambda_k). \tag{21}$$

Considering the terms which only depend on $\boldsymbol{\mu}_k$ in (20), $\ln q^*(\boldsymbol{\mu}_k|\lambda_k)$ is given by

$$\ln q^*(\boldsymbol{\mu}_k|\lambda_k) = -\lambda_k \boldsymbol{\mu}_k^H (\sum_{n=1}^{N} -\xi_{nk}\mathbf{x_n}\mathbf{x}_n^H + \beta_0\mathbf{I})\boldsymbol{\mu}_k + \lambda_k \boldsymbol{\mu}_k^H \beta_0 \mathbf{I}\mathbf{m_0} + \lambda_k \mathbf{m_0}^H \beta_0 \mathbf{I}\boldsymbol{\mu}_k + \text{const}, \tag{22}$$

where we have made use of $\boldsymbol{\mu}_k^H \boldsymbol{\mu}_k = 1$. Therefore, as a consequence of choosing a proper conjugate distribution, $q^*(\boldsymbol{\mu}_k|\lambda_k)$ is again a circular-symmetric complex-Gaussian distribution as

$$q^*(\boldsymbol{\mu}_k|\lambda_k) = \mathbb{N}_c\big(\boldsymbol{\mu}_k|\mathbf{m_k}, (\lambda_k\boldsymbol{\beta}_k)^{-1}\big), \tag{23}$$

where

$$\boldsymbol{\beta}_k = \sum_{n=1}^{N} -\xi_{nk}\mathbf{x_n}\mathbf{x_n}^H + \beta_0\mathbf{I} \tag{24}$$

$$\mathbf{m_k} = \boldsymbol{\beta}_k^{-1}\beta_0\mathbf{I}\mathbf{m}_0. \tag{25}$$

From (21), we can determine $q^*(\lambda_k)$ simply as

$$\ln q^*(\lambda_k) = \ln q^*(\boldsymbol{\mu}_k, \lambda_k) - \ln q^*(\boldsymbol{\mu}_k|\lambda_k). \tag{26}$$

On the right hand side of (26), we substitute for $\ln q^*(\boldsymbol{\mu}_k|\lambda_k)$ using (22) and for $\ln q^*(\boldsymbol{\mu}_k, \lambda_k)$ using (20). Keeping those terms that only depend on λ_k, we obtain

$$\ln q^*(\lambda_k) = \big((D-1)\sum_{n=1}^{N}\xi_{nk} + (a_0-1)\big)\ln\lambda_k - \lambda_k\big(\sum_{n=1}^{N}\xi_{nk}\mathbf{x}_n^H\mathbf{x}_n + \mathbf{m}_0^H\beta_0\mathbf{I}\mathbf{m}_0 + b_0 - \mathbf{m}_k^H\boldsymbol{\beta}_k\mathbf{m}_k\big) \tag{27}$$

Note that the terms involving $\boldsymbol{\mu}_k$ have been canceled out in (27) since $q^*(\lambda_k)$ is independent of $\boldsymbol{\mu}_k$. Hence, $q^*(\lambda_k)$ is a Gamma distribution as

$$q^*(\lambda_k) = \mathbb{G}(\lambda_k|a_k, b_k), \tag{28}$$

where shape parameter a_k and scale parameter b_k are given by

$$a_k = (D-1)\sum_{n=1}^{N}\xi_{nk} + a_0 \tag{29}$$

$$b_k = \sum_{n=1}^{N}\xi_{nk}\mathbf{x}_n^H\mathbf{x}_n + \mathbf{m}_0^H\beta_0\mathbf{I}\mathbf{m}_0 + b_0 - \mathbf{m}_k^H\boldsymbol{\beta}_k\mathbf{m}_k.^{1} \tag{30}$$

Finally, we showed that the posterior distribution $q^*(\boldsymbol{\mu}_k, \lambda_k)$, as expected, is a Gaussian-Gamma distribution, $q^*(\boldsymbol{\mu}_k, \lambda_k) = \mathbb{N}_c(\boldsymbol{\mu}_k|\mathbf{m_k}, (\boldsymbol{\beta}_k\lambda_k)^{-1})\mathbb{G}(\lambda_k|a_k, b_k)$, which is a consequence of using conjugate distributions.

In order to calculate the responsibilities ξ_{nk} (14), we need to calculate $\mathrm{E}_{\gamma_k}[\ln\gamma_k]$, $\mathrm{E}_{\lambda_k}[\ln\lambda_k]$, and $\mathrm{E}_{\boldsymbol{\mu}_k,\lambda_k}[\lambda_k\|\mathbf{x}_n - (\boldsymbol{\mu}_k^H\mathbf{x}_n)\boldsymbol{\mu}_k\|^2]$ which are involved in the calculation of (12). $\mathrm{E}_{\gamma_k}[\ln\gamma_k]$ and $\mathrm{E}_{\lambda_k}[\ln\lambda_k]$ are calculated by

$$\mathrm{E}_{\gamma_k}[\ln\gamma_k] = \digamma(\alpha_k) - \digamma(\hat{\alpha}), \tag{31}$$

$$\mathrm{E}_{\lambda_k}[\ln\lambda_k] = \digamma(a_k) - \ln b_k, \tag{32}$$

where $\digamma(\cdot)$ is the digamma function. For the calculation of $\mathrm{E}_{\boldsymbol{\mu}_k,\lambda_k}[\lambda_k\|\mathbf{x}_n - (\boldsymbol{\mu}_k^H\mathbf{x}_n)\boldsymbol{\mu}_k\|^2]$, first we calculate the required expectation with respect to $\boldsymbol{\mu}_k$, and then we take the expectation with respect to λ_k, which gives

$$\mathrm{E}_{\boldsymbol{\mu}_k,\lambda_k}[\lambda_k\|\mathbf{x}_n - (\boldsymbol{\mu}_k^H\mathbf{x}_n)\boldsymbol{\mu}_k\|^2] = \mathbf{x}_n^H\big((I - \mathbf{m}_k\mathbf{m}_k^H)\frac{a_k}{b_k} + D\boldsymbol{\beta}_k^{-1}\big)\mathbf{x}_n. \tag{33}$$

where we have made use of $\mathrm{E}[\boldsymbol{\mu}_k] = \mathbf{m}_k$, $\mathrm{E}[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^H] = \mathbf{m}_k\mathbf{m}_k^H + (\lambda_k\boldsymbol{\beta}_k)^{-1}$, and (21).

We can evaluate lower bound [7] for this model to monitor the bound during re-estimation and to examine the convergence. In every re-estimation, the value of the bound must increase. From [7], the lower bound can be calculated by

$$L_{\text{bound}} = \mathrm{E}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\lambda})] + \mathrm{E}[\ln p(\mathbf{Z}|\boldsymbol{\gamma})] + \mathrm{E}[\ln p(\boldsymbol{\gamma})] + \mathrm{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\lambda})] - \mathrm{E}[\ln q(\mathbf{Z})] - \mathrm{E}[\ln q(\boldsymbol{\gamma})] - \mathrm{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\lambda})], \tag{34}$$

where the expectation is taken with respect to the component parameters and latent variables, appropriately. Regarding to the model description and the derived update equations, we obtain

$$\begin{aligned} L_{\text{bound}} = \sum_{n=1}^{N}\sum_{k=1}^{K}\xi_{nk}\big(&-(D-1)\ln\pi + (D-1)\ln\mathrm{E}[\lambda_k] - \\ &\mathbf{x}_n^H((I - \mathbf{m}_k\mathbf{m}_k^H)\frac{a_k}{b_k} + D\boldsymbol{\beta}_k^{-1})\mathbf{x}_n\big) + \sum_{n=1}^{N}\sum_{k=1}^{K}\xi_{nk}\mathrm{E}[\ln\boldsymbol{\gamma}_k] \\ &+\ln C(\boldsymbol{\alpha}_0) + (\alpha_0 - 1)\sum_{k=1}^{K}\mathrm{E}[\ln\boldsymbol{\gamma}_k] + D\ln\frac{|\beta_0\mathbf{I}|}{\pi} + D\ln\lambda_k - \\ &D\beta_0\mathbf{I}\boldsymbol{\beta}_k^{-1} - \frac{a_k}{b_k}(\mathbf{m}_k - \mathbf{m}_0)^H(\mathbf{m}_k - \mathbf{m}_0) + \\ &\ln\frac{b_0^{a_0}}{\Gamma(a_0)} + (a_0 - 1)\mathrm{E}[\ln\lambda_k] - b_0\frac{a_k}{b_k} + \sum_{n=1}^{N}\sum_{k=1}^{K}\xi_{nk}\ln\xi_{nk} \\ &+\ln C(\boldsymbol{\alpha}_0) + \sum_{k=1}^{K}(\alpha_k - 1)\mathrm{E}[\ln\boldsymbol{\gamma}_k] + D\ln\mathrm{E}[\lambda_k] \\ &+D(\ln\boldsymbol{\beta}_k - \ln\pi - 1) - H[q(\lambda_k)], \end{aligned} \tag{35}$$

where $H[q(\lambda_k)]$ is the entropy of the Gamma distribution which is given by $H[q(\lambda_k)] = \ln\Gamma(\lambda_k) - (a-1)\digamma(\lambda_k) - \ln b_k + a_k$.

Fig. 1 shows the lower bound at a particular frequency bin, after applying the approach to a two-channel live-recording mixture of 3 speech signals in a room with 130 ms reverberation time and 1 m microphone spacing. The initial number of components K is set to 7. It is noticeable that there are four jumps in iterations 12, 53, 70, and 75. It implies that in those iterations, one of the components has been disregarded during the optimization procedure. More specifically, we demonstrate the classification result in Fig. 2 by two-dimensional plots of the real part of the projected posterior mean and the covariance in the mentioned iterations. By comparing Fig. 1 and Fig. 2, the jumps in Fig. 1 can be explained.

In summary, the algorithm starts with the initialization of the hyper-parameters characterizing the parameter distributions. In the next step, we use the current distribution over the model parameters to evaluate the responsibilities (14) by computing moments (31), (32), and (33) involved in the calculation of (12). Later, these responsibilities are used for the optimization of the variational posterior distribution over parameters using re-computing (18), (23), and (28). In every iteration, we monitor the variational lower bound (34), and the procedure will be continued until convergence. In other words, the optimization of the variational posterior distribution is analogous to the E and M step of the maximum likelihood EM algorithm.

[1] $\mathbf{m}_0^H\beta_0\mathbf{I}\mathbf{m}_0$ and $\mathbf{m}_k^H\boldsymbol{\beta}_k\mathbf{m}_k$ are real-valued scalars since $\beta_0\mathbf{I}$ and $\boldsymbol{\beta}_k$ are hermitan matrices.

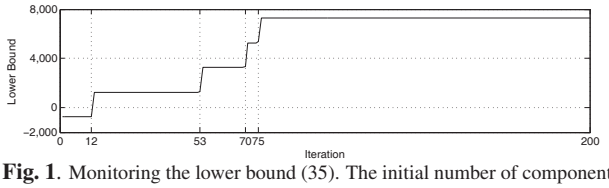

Fig. 1. Monitoring the lower bound (35). The initial number of components K is set to 7.

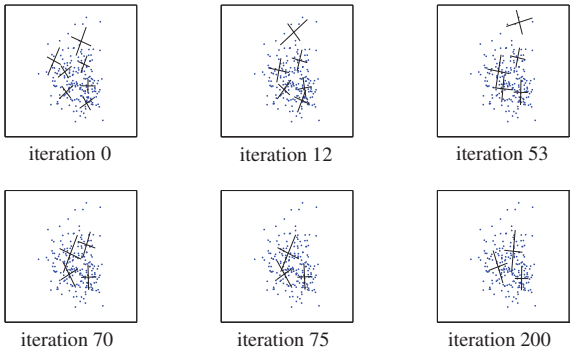

Fig. 2. Two-dimensional plots of the real part of the projected posterior mean and the covariance in iterations 12, 53, 70, 75, and 200.

5. SOURCE SEPARATION

There is a disorder along frequency bins, known as permutation ambiguity, so that the class order may differ from one frequency to another. Let $\boldsymbol{\xi}_j^f = [\xi_{1j}^f, \ldots, \xi_{Nj}^f]$ denote a sequence of responsibilities in a frequency bin f, where $j \in K_{opt}^f$, and $K_{opt}^f = \{k\}_{\alpha_k > \epsilon}$. ξ_{nk}^f and α_k are given by (14) and (19), respectively. The idea is that sequences belonging to the same source generally have similar patterns among different frequency bins. Therefore, the correlation coefficient of these sequences can be used to measure similarity among patterns and, hence, measure the interfrequency dependence. Here, we employ the permutation alignment procedure proposed in [1] for this purpose. There is an implementation issue which needs to be considered in the permutation alignment procedure. Let $\#K_{opt}^f$ denote the number of optimal components at frequency f. In the majority of frequencies, $\#K_{opt}^f$ is equal to the number of original sources Q, however, at some frequencies, $\#K_{opt}^f$ might be greater than or less than Q, which can be explained by the presence of noise in the environment or reverberation. Thus, we have three possible cases: $\#K_{opt}^f = Q$, $\#K_{opt}^f < Q$, and $\#K_{opt}^f > Q$. The second case is special case of the first case in the sense that one of the components is considered as zero. In the third case where $\#K_{opt}^f > Q$, we need to keep only Q of the components and disregard the other components. This is valid since α_k in those components are significantly less than the other components and most likely those components are related to the noise or reverberation.In this work, we use binary masking for constructing separated signals in the frequency domain. Binary masking relies on the sparseness property of the speech signal. Based on the sparsity property, at most one source has a large contribution to each TF point. The separated signals in the TF domain are constructed by $\hat{y}_{qd}(\tau, f) = \mathcal{M}_k(\tau, f) y_d(\tau, f)$, where we have defined $\mathcal{M}_k(\tau, f)$ such that $\mathcal{M}_k(\tau, f) = 1$ if $\xi_{nk} \geq \xi_{nk'}$ and $\mathcal{M}_k(\tau, f) = 0$ if $\xi_{nk} < \xi_{nk'}, \forall k' \neq k$. Finally, separated signals $\hat{y}_{qd}(\tau, f)$ are transformed to the time domain by the inverse short-time Fourier transform.

6. EXPERIMENTAL RESULTS

The algorithm is evaluated on the development data (dev1.zip) used in the audio source separation campaign (SiSEC08) [9]. We consider live recording mixtures of three female speech signals (female3) and 3 male speech signals (male3) sampled at 16 kHz and with 9 seconds duration. In order to evaluate the robustness of the proposed algorithm, two microphone spacings, 5 cm and 1 m, are considered under 130 ms and 250 ms reverberation times. The algorithm uses a 2048 sample length Hann window with a 75% overlap. Table 1 shows the separation results in terms of the output signal-to-distortion ratio (SDR) [10] for the proposed variational Bayes approach and EM-based approach [1]. This table shows the best and average results of 20 times running the algorithms. The EM based algorithm benefits from the prior knowledge on the exact number of sources (the number of components K is set to the number of sources), however, the variational Bayes approach assumes no prior information on the number of sources (K is set to an arbitrary large value, i.e., $K = 7$ in our experiments). Although EM-based approach has slightly better performance, it requires a prior knowledge on the number of sources while the proposed variational Bayes approach does not. The audio files of the experiments can be downloaded from [11]. We also examined the algorithm for scenarios where more than three sources are mixed. However, we run into some instability, that we think could possibly be related either to the permutation alignment procedure or the classification part. This needs to be examined in details in the future work.

Table 1. Separation results on SiSEC 2008 database in terms of the average output SDR of all sources in dB. VB refers to the proposed variational Bayes approach and EM refers to the EM-based approach. SDR* and $\overline{\text{SDR}}$ show, respectively, the best and the average results in 20 times running both algorithms with random initializations

RT_{60}		130ms		250ms		130ms		250ms	
mixture		female3		female3		male3		male3	
mic. spacing		1m	5cm	1m	5cm	1m	5cm	1m	5cm
VB	SDR*	7.7	5.8	6.7	5.5	6.4	5.3	6.0	4.6
	$\overline{\text{SDR}}$	6.2	5.0	5.7	4.5	5.2	4.2	5.6	3.0
EM	SDR*	8.6	6.4	7.3	5.8	6.7	5.8	6.1	4.3
	$\overline{\text{SDR}}$	6.8	5.5	6.4	5.0	5.7	4.6	5.8	3.6

7. CONCLUSION

We proposed a variational Bayes approach to the underdetermined blind source separation in the time-frequency domain. We showed that the proposed approach has a competitive separation performance compared to the EM-based approach [1] while it requires no prior knowledge on the number of sources and with little overhead in terms of the computational complexity.

8. REFERENCES

[1] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516 –527, 2011.

[2] P. D. O'Grady and B. A. Pearlmutter, "The lost algorithm: Finding lines and separating speechmixtures," *EURASIP J. A. Signal Process.*, vol. Article ID 784296, 2008.

[3] N. Nasios and A.G. Bors, "Variational learning for gaussian mixture models," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 4, pp. 849 –862, Aug. 2006.

[4] M.I. Mandel, R.J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 18, no. 2, pp. 382 –394, feb. 2010.

[5] Ono N. Izumi, Y. and S. Sagayama, "Sparseness-based 2ch bss using the em algorithm in reverberant environment," in *Proc. WASPAA*, 2007.

[6] H. S. Stern D. B. Rubin A. Gelman, J. B. Carlin, *Bayesian Data Analysis*, Chapman & Hall, 1995.

[7] C. M. Bshop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[8] J. Karhunen A. Hyvrinen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[9] www.sisec2008.wiki.irisa.fr/tiki index.php.

[10] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. ICA*, 2007, pp. 552 –559.

[11] www.ee.kth.se/∼taghia/ICASSP2012.zip.