Nonstationary Source Separation Using Sequential and Variational Bayesian Learning

Jen-Tzung Chien, *Senior Member, IEEE*, and Hsin-Lung Hsieh

***Abstract*— Independent component analysis (ICA) is a popular approach for blind source separation where the mixing process is assumed to be unchanged with a fixed set of stationary source signals. However, the mixing system and source signals are nonstationary in real-world applications, e.g., the source signals may abruptly appear or disappear, the sources may be replaced by new ones or even moving by time. This paper presents an online learning algorithm for the Gaussian process (GP) and establishes a separation procedure in the presence of nonstationary and temporally correlated mixing coefficients and source signals. In this procedure, we capture the evolved statistics from sequential signals according to online Bayesian learning. The activity of nonstationary sources is reflected by an automatic relevance determination, which is incrementally estimated at each frame and continuously propagated to the next frame. We employ the GP to characterize the temporal structures of time-varying mixing coefficients and source signals. A variational Bayesian inference is developed to approximate the true posterior for estimating the nonstationary ICA parameters and for characterizing the activity of latent sources. The differences between this ICA method and the sequential Monte Carlo ICA are illustrated. In the experiments, the proposed algorithm outperforms the other ICA methods for the separation of audio signals in the presence of different nonstationary scenarios.**

***Index Terms*— Bayes procedure, blind source separation (BSS), Gaussian process (GP), independent component analysis (ICA), online learning, variational method.**

I. Introduction

BLIND source separation (BSS) attempts to recover the independent source signals $\mathbf{s}_t = \left[s_{1,t}, \ldots, s_{M,t}\right]^T$ at frame t under the situations that we only observe the mixed signals $\mathbf{x}_t = \left[x_{1,t}, \ldots, x_{N,t}\right]^T$, and the actual mixing process, expressed by $\mathbf{x}_t = \mathbf{A}\mathbf{s}_t$ with a $N \times M$ mixing matrix $\mathbf{A}$, is unknown. The source signals are assumed to be stationary and determined by $\mathbf{s}_t = \mathbf{W}\mathbf{x}_t$ where the demixing matrix $\mathbf{W}$ or the mixing matrix $\mathbf{A}$ is also assumed to be stationary and can be estimated by an independent component analysis (ICA) procedure. The conventional ICA methods were developed by maximizing the kurtosis or minimizing the mutual information between demixed signals [3], [10], and [13]. A BSS algorithm [46] was established to find the demixing matrix by maximizing the temporal predictability where the covariances between signal mixtures are computed. More recently, BSS methods [4], [36], [47], and [49] were developed by conducting matrix factorization based on the time-frequency analysis [50] or using the regularized priors [20], [21]. The single-channel BSS problem was also extensively studied [14], [19], [42]. In addition, ICA was employed to conduct independent vector analysis for joint BSS over multiple datasets [2]. In [51], a sparse component analysis was introduced to deal with the nonnegative BSS where the demixing matrix was estimated via a quadratic programming technique. All of these methods did not consider the nonstationary conditions, the effect of noise signal, and the uncertainty of parameters in the ICA or BSS generative model.

In real-world applications, the mixing system may involve various nonstationary scenarios due to the moving sources, the sudden presence or absence of sources, or even the original source being replaced by a new one. Since the mixing coefficients are affected by the distance between source and sensor, the nonstationary source separation turns out to deal with the nonstationary source signals and mixing coefficients. To solve such a complicated circumstance, we may detect the status of source signals and adapt the source distributions at each frame. Several methods have been proposed to cope with dynamic sources and a nonstationary mixing system by investigating the following two scenarios separately. First, the sources or sensors are moving. For this scenario, an adaptive BSS algorithm was used to compensate the variations of a mixing matrix. A Markov process was applied to capture the variations by tracking the mixtures of temporally correlated sources [18]. Also, a 3-D tracker was used to detect the status of sources [38]. If the sources were moving, a beamforming algorithm was applied for BSS. The source distributions and the number of sources were assumed to be fixed. In the second scenario, the sources may suddenly appear, disappear or even be replaced by new ones at different frames. The time-varying source distributions should be characterized. The scheme of automatic relevance determination (ARD) [25], [33], [48] was employed to reflect the activity of source signals. The abrupt presence and absence of sources was captured by an indicator variable in a switching ICA (S-ICA) [23]. A hidden Markov model (HMM) was constructed to represent the status of the source signals while the generative model was assumed to be fixed. The replacement of source signals was tackled [15]. Further, an online variational Bayesian (VB) learning [24] was performed in an ICA procedure where the

Manuscript received May 10, 2012; revised January 13, 2013; accepted January 16, 2013. Date of publication February 4, 2013; date of current version March 8, 2013. This work was supported in part by the National Science Council, Taiwan, under Contract NSC 100-2628-E-009-028-MY3.

The authors are with the Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: jtchien@nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2013.2242090

source signal was a time-dependent parameter but the mixing matrix was time independent. The source distributions were updated incrementally with a fixed number of sources by disregarding the sudden presence and absence of sources. A non-Gaussianity-based piecewise stationary ICA [30] was proposed to explore the varying distribution of non-Gaussian signals for the separation of audio signals.

This paper presents online Bayesian learning for nonstationary source separation in which two scenarios were investigated simultaneously [25], [26]. Unlike the sequential Monte Carlo ICA [1], [6] based on sequential importance sampling procedure, we develop a new online learning algorithm by characterizing the temporally correlated mixing coefficients and source signals. The ICA model is constructed by marginalizing the uncertainty of model parameters via VB inference [8]. We propose an ICA algorithm based on online GP and apply it to dynamic source separation. The temporally correlated mixing coefficients and source signals are characterized by GP where the posterior statistics are accumulated and propagated frame-by-frame according to a recursive Bayesian algorithm [9], [11], [12], [45]. The GP prior is merged to represent the temporal structures of the mixing process and source signal. The online Bayesian learning is performed by combining the prior distribution updated from previous frames and the likelihood function calculated by using the current frame. The updated posteriors act as new priors for the prediction of the next frame. A compensation parameter is introduced to adjust the distribution of the mixing matrix for an ARD scheme. The temporally correlated mixing coefficients and source signals are estimated by maximizing the marginal likelihood so as to achieve the highest temporal predictability. The proposed ICA is investigated by the experiments on source separation of audio signals under different nonstationary scenarios.

The remainder of this paper is organized as follows. In the next section, we survey several source separation methods. The proposed methods are overviewed. Section III presents the online GP for nonstationary source separation in the presence of temporally correlated mixing coefficients and source signals. Section IV addresses the sequential and variational inference for ICA methods. Section V reports a series of experiments on separation of speech and music signals in nonstationary environments. The conclusions drawn from this paper are given in Section VI.

II. Nonstationary Source Separation

Standard ICA assumes that the observations are mixed by a fixed set of independently and identically distributed sources. No noise signal is considered. These assumptions are not realistic in nonstationary environments. Two categories of nonstationary source separation methods are surveyed.

A. Separation Based on a Nonstationary Source Model

In many applications, the status of source signals is changed at different time frames. A nonstationary model was built to meet the changing distributions of source signals. An HMM was incorporated to catch the temporal information of source signals according to a noisy ICA model [15], [29], [43]

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \boldsymbol{\varepsilon}_t \tag{1}$$

with a noise signal $\boldsymbol{\varepsilon}_t$. The mth source signal $s_{m,t}$ was modeled by a mixture of Gaussians (MoG) in

$$p(\mathbf{s}_t|\boldsymbol{\Theta}) = \prod_{m=1}^{M}\left[\sum_{k=1}^{K}\pi_{mk}\mathcal{N}(s_{m,t}|\mu_{mk}, \gamma_{mk}^{-1})\right] \tag{2}$$

where $\boldsymbol{\Theta} = \{\pi_{mk}, \mu_{mk}, \gamma_{mk}\}$ denotes the state-dependent Gaussian mixture parameters consisting of mixture weights $\{\pi_{mk}\}$, means $\{\mu_{mk}\}$, and precisions $\{\gamma_{mk}\}$ from K Gaussian components. Equation (2) was obtained due to M mutually independent sources. The temporal information of source signals was characterized by HMM states. This method captured the nonstationary source variables via VB learning.

To relax the assumption of a fixed set of source signals at different time frames, a S-ICA [23] was proposed to overcome the circumstance that source signal $s_{m,t}$ was dynamically active or inactive. The status was indicated by a switching variable with $z_{m,t} = 0$ indicating an inactive source and $z_{m,t} = 1$ indicating an active source. The source signal in state $z_{m,t} = 1$ was modeled by $p(s_{m,t}|z_{m,t} = 1)$ using

$$\pi_m\mathcal{N}(s_{m,t}|0, \gamma_{ma}^{-1}) + (1 - \pi_m)\mathcal{N}(s_{m,t}|0, \gamma_{mb}^{-1}) \tag{3}$$

where π_m denotes the mixture weight and $\{\gamma_{ma}, \gamma_{mb}\}$ denotes the precisions of two Gaussians. The source signal was set to be zero when switching to state $z_{m,t} = 0$. A Markov process was used to represent the dynamics of switching variable. The initial state probability $p(z_{m,1})$ and the state transition probability $p(z_{m,t}|z_{m,t-1})$ were estimated. The key idea of S-ICA was to identify the Markov state of switching variable. The computation highly depended on the number of states, which was expanded exponentially by the number of sources M.

B. Separation Based on Temporal Structure

The sources from speaker and music signals are temporally correlated. The correlation information is crucial for signal reconstruction in nonstationary environments. The autoregressive (AR) process [22], [27], [28], and the GP [39] were used to discover temporal structure of source signals for nonstationary source separation. Using an AR process, a new sample $s_{m,t}$ at time t was predicted by using its past p samples $\vec{\mathbf{s}}_{m,t-1} = [s_{m,t-1}, \ldots, s_{m,t-p}]^T$ via a latent function

$$f(\vec{\mathbf{s}}_{m,t-1}) = \sum_{\tau=1}^{p} h_{m,\tau}s_{m,t-\tau} \tag{4}$$

where $h_{m,\tau}$ denotes the AR coefficients. In contrast with AR prediction using a linear parametric model, GP was employed to build a nonlinear nonparametric regression model where $s_{m,t}$ was predicted according to a zero-mean Gaussian prior by using a kernel-function-based covariance parameter [39]

$$\mathcal{N}(f(\vec{\mathbf{s}}_{m,t-1})|0, \kappa(\vec{\mathbf{s}}_{m,t-1}, \vec{\mathbf{s}}_{m,\tau-1})) \tag{5}$$

where $f(\cdot)$ denotes a latent function for GP. Any subset of source signals has a joint Gaussian distribution. The temporal

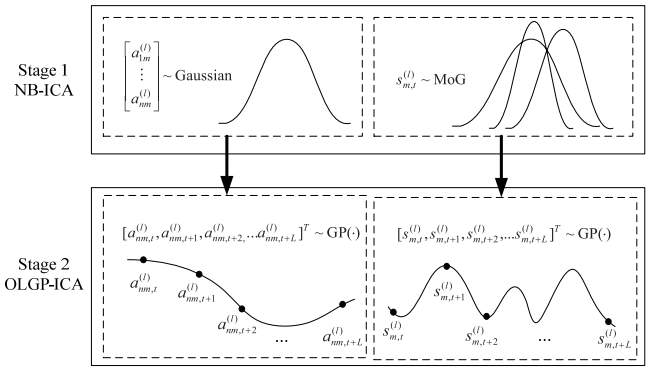

Fig. 1. Evolution of ICA methods for nonstationary source separation.

structure of the mth source was characterized by a regression function

$$s_{m,t} = f(\vec{\mathbf{s}}_{m,t-1}) + \varepsilon_{m,t} \tag{6}$$

where $\varepsilon_{m,t}$ denotes a white Gaussian noise with zero mean and unit variance. The GP model was flexible by working on a high dimensional kernel space with a $L \times L$ positive definite Gram matrix $\mathbf{K}_{s_m}$ given by [40]

$$[\mathbf{K}_{s_m}]_{t\tau} = \kappa(\vec{\boldsymbol{s}}_{m,t-1}, \vec{\boldsymbol{s}}_{m,\tau-1}) + \delta_{t\tau} \tag{7}$$

where $\delta_{t\tau} = 1$ at $t = \tau$ and $\delta_{t\tau} = 0$ otherwise.

C. Overview of Proposed Methods

The previous methods are developed for batch learning where the model parameters are estimated from a batch data collection with L samples. Although the nonstationary source signals were characterized by Markov chain [15], [23], [29], [43], AR [27], [28], and GP [39], a single set of model parameters is insufficient to elaborately compensate the temporally correlated source signals under nonstationary environments. Detecting the status of source signals and adaptively changing their distributions at each frame are useful approaches. Besides, the real-world mixing system may contain active or inactive sources, the moving sources or even the replacement of a new source. Considering the temporally correlated mixing coefficients is beneficial to improve BSS performance. In this paper, we evolve the nonstationary source separation methods by two stages as shown in Fig. 1. First of all, the mixing coefficients are assumed to be distributed by a fixed Gaussian distribution within a frame, but the distribution is varied frame-by-frame. The source signal is distributed by an MoG. An online Bayesian learning procedure is presented to continuously update the variational posteriors and their hyperparameters. A VB procedure is established to fulfill this nonstationary Bayesian ICA (denoted by NB-ICA) algorithm. In the second stage, the mixing coefficients are not only distributed differently across frames but also treated as temporally correlated variables within a frame. We tackle the temporally correlated mixing coefficients and source signals for dynamic source separation. The temporal structures are characterized by GP. The ICA algorithm based on online GP (denoted by OLGP-ICA) is developed through a VB procedure.

III. Online Learning and the Gaussian Process

A. Online Bayesian Learning

Bayesian approaches are important to build the regularized model and avoid the overfitting problem [5]. Bayesian ensemble learning for ICA has been studied in [16], [34], [35], and [41]. In general, conventional ICA methods assumed that the source signals were independently and identically distributed, and the mixed signals were generated from a fixed mixing system. These assumptions are not fitted to the nonstationary conditions. A second-order Markov process was proposed for nonstationary ICA [34]. Also, the online learning procedure can be applied to solve nonstationary BSS by incrementally estimating the dynamic sources via the ICA algorithm [1], [6], [25], [26]. Using the recursive Bayesian learning [9], [12], [41], [45], the sufficient statistics from previous frames are combined with the likelihood of the current frame. The resulting posterior distribution is then propagated to the next learning epoch for adaptive BSS. The distributions of source signals and mixing coefficients are adapted to meet the changing environments at different frames. Let $\mathbf{X}^{(l)} = \{\mathbf{x}_t^{(l)}\}$ denote a set of mixed signals at frame l with L samples. The signals are mixed by a linear combination of M unknown source signals $\mathbf{S}^{(l)} = \{\mathbf{s}_t^{(l)}\}$ using a mixing matrix $\mathbf{A}^{(l)}$

$$\mathbf{x}_t^{(l)} = \mathbf{A}^{(l)}\mathbf{s}_t^{(l)} + \boldsymbol{\varepsilon}_t^{(l)} \tag{8}$$

where $\mathbf{E}^{(l)} = \{\boldsymbol{\varepsilon}_t^{(l)}\}$ denotes the noise signals. The distribution and the activity of sources are assumed to be unchanged within a frame but varied across frames. We attempt to incrementally characterize the variations of $\mathbf{A}^{(l)}$ and $\mathbf{S}^{(l)}$ from the observed frames $\mathcal{X}^{(l)} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(l)}\}$ through the stages of *prediction* and *correction*. First, the ICA model parameters $\boldsymbol{\Theta}^{(l)}$ at frame l are predicted according to the posterior distribution given the previous frames $\mathcal{X}^{(l-1)}$ [18]

$$\begin{aligned} &p(\boldsymbol{\Theta}^{(l)}|\mathcal{X}^{(l-1)}) \\ &= \int p(\boldsymbol{\Theta}^{(l)}|\boldsymbol{\Theta}^{(l-1)})\, p(\boldsymbol{\Theta}^{(l-1)}|\mathcal{X}^{(l-1)})\, d\boldsymbol{\Theta}^{(l-1)} \end{aligned} \tag{9}$$

which is obtained by integrating over the uncertainty of previous parameters $\boldsymbol{\Theta}^{(l-1)}$. Equation (9) is known as the predictive distribution, which is essential in full Bayesian framework [5]. The prediction stage is performed in a VB procedure for ICA model inference, which will be addressed in Section IV. Optimizing the predictive distribution is fulfilled to establish a noisy ICA model of (8).

Second, when a new frame with the mixed samples $\mathbf{X}^{(l)} = \{\mathbf{x}_t^{(l)}\}$ is observed, the posterior distribution is corrected by [45]

$$p(\boldsymbol{\Theta}^{(l)}|\mathcal{X}^{(l)}) = \frac{p(\mathbf{X}^{(l)}|\boldsymbol{\Theta}^{(l)})\, p(\boldsymbol{\Theta}^{(l)}|\mathcal{X}^{(l-1)})}{\int p(\mathbf{X}^{(l)}|\boldsymbol{\Theta}^{(l)})\, p(\boldsymbol{\Theta}^{(l)}|\mathcal{X}^{(l-1)})\, d\boldsymbol{\Theta}^{(l)}} \tag{10}$$

which is proportional to the product of a likelihood function of current frame $\mathbf{X}^{(l)}$ and *a posteriori* distribution given the previous frames $\mathcal{X}^{(l-1)}$. At each learning epoch l, the posterior distribution $p(\boldsymbol{\Theta}^{(l)}|\mathcal{X}^{(l-1)})$ is seen as a prior distribution $p(\boldsymbol{\Theta}^{(l)}|\boldsymbol{\Phi}^{(l-1)})$ with hyperparameters $\boldsymbol{\Phi}^{(l-1)}$, which are updated from the previous data $\mathcal{X}^{(l-1)}$. We choose the

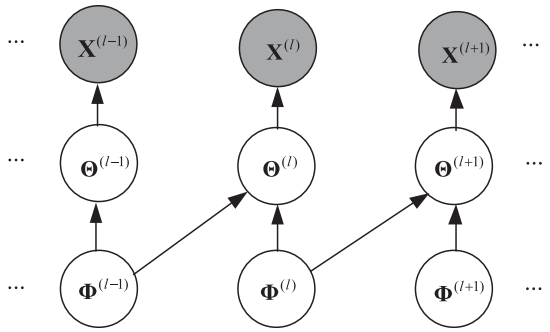

Fig. 2. Three layers of variables in online Bayesian learning.

conjugate prior so that the updated posterior $p(\boldsymbol{\Theta}^{(l)}|\mathcal{X}^{(l)})$ has the same distribution form as its prior density $p(\boldsymbol{\Theta}^{(l)}|\boldsymbol{\Phi}^{(l-1)})$. The reproducible prior/posterior distribution pair is formed for incremental learning of the posterior distribution or the prior density with hyperparameters $\boldsymbol{\Phi}^{(0)} \rightarrow \boldsymbol{\Phi}^{(1)} \rightarrow \ldots \rightarrow \boldsymbol{\Phi}^{(l)}$. With the updated hyperparameters, the ICA parameters correspond to the modes of posterior distribution and can be realized by an online manner $\boldsymbol{\Theta}^{(0)} \rightarrow \boldsymbol{\Theta}^{(1)} \rightarrow \ldots \rightarrow \boldsymbol{\Theta}^{(l)}$. The newest environments are characterized for nonstationary source separation. A recursive Bayesian algorithm with three layers of variables is depicted in Fig. 2. Unlike batch learning [15], [23], we present an online learning approach to detect the activities of source signals and estimate the distributions of reconstructed sources frame-by-frame. After updating the hyperparameters, the current frame $\mathbf{X}^{(l)}$ is abandoned and only the sufficient statistics $\boldsymbol{\Phi}^{(l)}$ are stored and propagated to the next learning epoch $l+1$.

B. Nonstationary Bayesian ICA

First of all, we conduct online learning without considering the temporal structures of the mixing process and source model. The distributions of mixing matrix $\mathbf{A}^{(l)}$ and source signals $\mathbf{s}_t^{(l)}$ are fixed within a frame l. The NB-ICA algorithm [25] is proposed. Considering the noisy ICA model in (8), the source signals $\mathbf{s}_t^{(l)}$ are distributed by a MoG with K Gaussians

$$p(\mathbf{s}_t^{(l)}|\boldsymbol{\Pi}^{(l)} = \{\pi_{mk}^{(l)}\}, \mathbf{M}^{(l)} = \{\mu_{mk}^{(l)}\}, \mathbf{R}^{(l)} = \{\gamma_{mk}^{(l)}\}) = \prod_{m=1}^{M}\left[\sum_{k=1}^{K}\pi_{mk}^{(l)}\mathcal{N}(s_{m,t}^{(l)}|\mu_{mk}^{(l)}, (\gamma_{mk}^{(l)})^{-1})\right]. \quad (11)$$

The noise vector $\boldsymbol{\varepsilon}_t^{(l)}$ is assumed to be Gaussian $\mathcal{N}(\varepsilon_t^{(l)}|0, (\mathbf{B}^{(l)})^{-1})$ with zero mean and diagonal precision matrix $\mathbf{B}^{(l)} = \text{diag}\{\beta_n^{(l)}\}$. The resulting likelihood function of observation frame $\mathbf{x}_t^{(l)}$ is written by

$$p(\mathbf{x}_t^{(l)}|\mathbf{A}^{(l)}, \mathbf{s}_t^{(l)}, \boldsymbol{\varepsilon}_t^{(l)}, \mathbf{B}^{(l)}) = \mathcal{N}(\mathbf{x}_t^{(l)}|\mathbf{A}^{(l)}\mathbf{s}_t^{(l)}, (\mathbf{B}^{(l)})^{-1}). \quad (12)$$

The prior density of $N \times M$ mixing matrix $\mathbf{A}^{(l)} = \{a_{nm}^{(l)}\}$ is distributed by

$$p(\mathbf{A}^{(l)}|\boldsymbol{\alpha}^{(l)}) = \prod_{n=1}^{N}\left[\prod_{m=1}^{M}\mathcal{N}(a_{nm}^{(l)}|0, (\alpha_m^{(l)})^{-1})\right] = \prod_{m=1}^{M}\mathcal{N}(\mathbf{a}_m^{(l)}|0, (\alpha_m^{(l)})^{-1}\mathbf{I}_N) \quad (13)$$

where $\boldsymbol{\alpha}^{(l)} = \{\alpha_m^{(l)}\}$ and $\mathbf{I}_N$ is a N-dimensional identity matrix. Each column $\mathbf{a}_m^{(l)}$ of $\mathbf{A}^{(l)}$ has an isotropic Gaussian distribution with zero mean and precision $\alpha_m^{(l)}$. Importantly, if the precision $\alpha_m^{(l)}$ in (13) is gamma distributed, the marginal distribution of mixing coefficient $a_{nm}^{(l)}$ over gamma prior of $\alpha_m^{(l)}$ turns out to be a Student's t distribution, which is peaky around zero and is popular for sparse Bayesian learning [48]. The hyperparameter $\alpha_m^{(l)}$ is known as an ARD [33], [48], which reveals the activity of a source signal $s_{m,t}^{(l)}$ in ICA model. The matrix $\mathbf{A}^{(l)}$ is prone to be sparse with near zero entries at the mth column $\mathbf{a}_m^{(l)} = \{a_{nm}^{(l)}\} \rightarrow 0$ if the estimated ARD $\alpha_m^{(l)}$ is large. The mth source is likely inactive at frame l. The redundant sources are disregarded automatically. This ARD $\alpha_m^{(l)}$ is similar to the indicator variable $z_{m,t}$ in the S-ICA [23].

However, each coefficient $a_{nm}^{(l)}$ reflects the mixing correlation between source m and sensor n, which is continuously varied in nonstationary environments. A single ARD parameter $\alpha_m^{(l)}$ is not sufficient to indicate the relevance between source m and N different sensors. To compensate this weakness, the precision matrix of prior density of $\mathbf{a}_m^{(l)}$ is adapted by using a transformation matrix $\mathbf{H}_m^{(l)}$. The prior density of $\mathbf{A}^{(l)}$ in (13) is modified as

$$p(\mathbf{A}^{(l)}|\boldsymbol{\alpha}^{(l)}, \mathbf{H}^{(l)}) = \prod_{m=1}^{M}\mathcal{N}(\mathbf{a}_m^{(l)}|0, (\alpha_m^{(l)}\mathbf{H}_m^{(l)})^{-1}) \quad (14)$$

where $\mathbf{H}^{(l)} = \{\mathbf{H}_m^{(l)}\}$. Figure 3 displays the graphical representation of NB-ICA model. The parameter set is formed by $\boldsymbol{\Theta}^{(l)} = \{\mathbf{A}^{(l)}, \boldsymbol{\alpha}^{(l)}\mathbf{H}^{(l)}, \mathbf{E}^{(l)}, \mathbf{B}^{(l)}, \mathbf{S}^{(l)}, \boldsymbol{\Pi}^{(l)}, \mathbf{M}^{(l)}, \mathbf{R}^{(l)}\}$. The online Bayesian learning is developed by specifying the conjugate priors for individual components of $\boldsymbol{\Pi}^{(l)}$, $\mathbf{M}^{(l)}$ and $\mathbf{R}^{(l)}$, which are Dirichlet, Gaussian, and gamma distributions with hyperparameters $\boldsymbol{\Phi}_\pi^{(l-1)}$, $\boldsymbol{\Phi}_m^{(l-1)}$, and $\boldsymbol{\Phi}_r^{(l-1)}$, respectively. The precision parameter $\beta_n^{(l)}$ is generated by a gamma prior with hyperparameters $\{u_{\beta_n}^{(l-1)}, \omega_{\beta_n}^{(l-1)}\}$. The new hyperparameters $\boldsymbol{\Phi}^{(l)}$ are estimated from current frame $\mathbf{X}^{(l)}$ and previous hyperparameters $\boldsymbol{\Phi}^{(l-1)}$. The detailed solution to $\boldsymbol{\Phi}^{(l)} = \{\mathbf{u}^{(l)} = \{u_{\beta_n}^{(l)}\}, \boldsymbol{\omega}^{(l)} = \{\omega_{\beta_n}^{(l)}\}, \boldsymbol{\Phi}_\pi^{(l)}, \boldsymbol{\Phi}_m^{(l)}, \boldsymbol{\Phi}_r^{(l)}\}$ has been derived in [16]. Nevertheless, the proposed NB-ICA has twofold novelties. One is the online learning and the other is the compensation of precision matrix of $\mathbf{a}_m^{(l)}$ by $\alpha_m^{(l)}\mathbf{I}_N \rightarrow \alpha_m^{(l)}\mathbf{H}_m^{(l)}$. A Wishart distribution is used as the conjugate prior to characterize the ARD parameter by

$$p(\alpha_m^{(l)}\mathbf{H}_m^{(l)}|\rho_m^{(l-1)}, \mathbf{V}_m^{(l-1)}) \propto |\alpha_m^{(l)}\mathbf{H}_m^{(l)}|^{(\rho_m^{(l-1)}-N-1)/2} \times \exp\left[-\frac{1}{2}\text{Tr}[(\mathbf{V}_m^{(l-1)})^{-1}\alpha_m^{(l)}\mathbf{H}_m^{(l)}]\right]. \quad (15)$$

The hyperparameters $\boldsymbol{\Phi}^{(l-1)} = \{\boldsymbol{\rho}^{(l-1)} = \{\rho_m^{(l-1)}\}, \mathbf{V}^{(l-1)} = \{\mathbf{V}_m^{(l-1)}\}\}$ from previous data $\mathcal{X}^{(l-1)}$ are applied. The solution to hyperparameters $\boldsymbol{\Phi}^{(l)} = \{\boldsymbol{\rho}^{(l)}, \mathbf{V}^{(l)}\}$ should be formulated for NB-ICA. Notably, the marginal distribution of mixing parameter $\mathbf{a}_m^{(l)}$ over Wishart prior of $\alpha_m^{(l)}\mathbf{H}_m^{(l)}$ is formed by a multivariate Student's t distribution. The nth diagonal entry of the precision matrix $\alpha_m^{(l)}\mathbf{H}_m^{(l)}$ reveals the relevance information about source m appearing in sensor n.

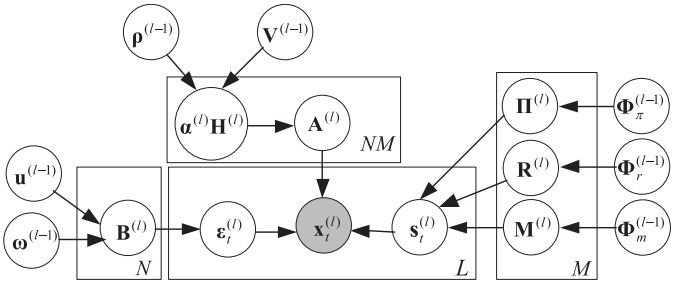

Fig. 3. Graphical representation of the NB-ICA model.

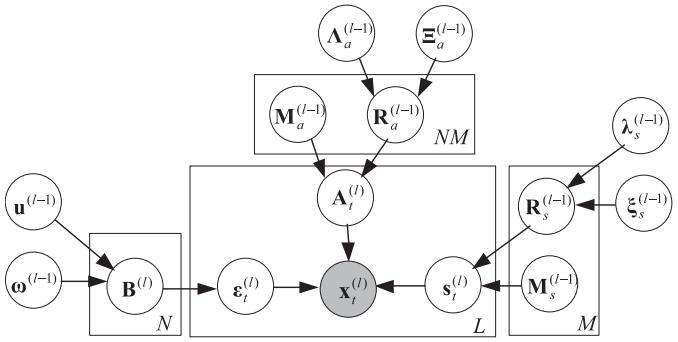

Fig. 4. Graphical representation of the OLGP-ICA model.

C. OLGP for ICA

Next, online learning algorithm is introduced by combining the temporally correlated mixing coefficients and source signals. As we know, the mixing coefficient $a_{nm}^{(l)}$ represents the mixing relation between source m and sensor n. The mixing matrix is time-varying $\mathbf{A}^{(l)} \rightarrow \mathbf{A}_t^{(l)}$ in presence of moving sources or sensors. Correspondingly, the mixing coefficients $\{a_{nm,t}^{(l)}\}$ of L samples at frame l are temporally correlated and their temporal structure is continuously changed across frames [26]. A flexible ICA model should be established by simultaneously performing the online learning and exploring the temporal structure of mixing coefficients and source signals. This paper presents an ICA algorithm based on online GP. The online Bayesian learning is undertaken to update system parameters $\mathbf{\Theta}^{(l)}$ or hyperparameters $\mathbf{\Phi}^{(l)}$ frame-by-frame. Unlike NB-ICA using a fixed $\mathbf{A}^{(l)}$, OLGP-ICA applies the time-varying mixing matrix $\mathbf{A}_t^{(l)}$. The temporal structure in L samples of $\{a_{nm,t}^{(l)}\}$ and $\{s_{m,t}^{(l)}\}$ is represented by GP and is merged in online Bayesian learning.

Model construction is addressed as follows. The noisy ICA model with time-varying mixing matrix $\mathbf{A}_t^{(l)}$ is considered. The temporally correlated mixing coefficients and source signals are generated by the distributions of nonparametric latent functions. Regarding the mixing coefficient, GP could flexibly explore the unknown temporal structure of $a_{nm,t}^{(l)}$. A latent function $f(\cdot)$ is employed to connect the relation between current coefficient $a_{nm,t}^{(l)}$ and its past p coefficients $\vec{\mathbf{a}}_{nm,t-1}^{(l)} = [a_{nm,t-1}^{(l)}, \ldots, a_{nm,t-p}^{(l)}]^T$ by

$$a_{nm,t}^{(l)} = f(\vec{\mathbf{a}}_{nm,t-1}^{(l)}) + \varepsilon_{nm,t}^{(l)} \tag{16}$$

where $\varepsilon_{nm,t}^{(l)}$ denotes the white noise. This function is generated from a zero-mean Gaussian

$$\mathcal{N}(f(\vec{\mathbf{a}}_{nm,t-1}^{(l)})|0, \kappa(\vec{\mathbf{a}}_{nm,t-1}^{(l)}, \vec{\mathbf{a}}_{nm,\tau-1}^{(l)})) \tag{17}$$

with a variance $\kappa(\vec{\mathbf{a}}_{nm,t-1}^{(l)}, \vec{\mathbf{a}}_{nm,\tau-1}^{(l)})$ given by

$$\xi_{a_{nm}}^{(l-1)} \exp\left[-\frac{\lambda_{a_{nm}}^{(l-1)}}{2}\left\|\vec{\mathbf{a}}_{nm,t-1}^{(l)} - \vec{\mathbf{a}}_{nm,\tau-1}^{(l)}\right\|^2\right] \tag{18}$$

which is an exponential-quadratic kernel function with parameters $\{\lambda_{a_{nm}}^{(l-1)}, \xi_{a_{nm}}^{(l-1)}\}$. Therefore, the Gaussian prior over a set of latent functions $\{f(\vec{\mathbf{a}}_{nm,t-1}^{(l)})\}$ at frame l is used to determine the GP prior $p(\mathbf{a}_{nm}^{(l)}|\boldsymbol{\mu}_{a_{nm}}^{(l-1)}, \mathbf{R}_{a_{nm}}^{(l-1)})$ for the mixing coefficients $\mathbf{a}_{nm}^{(l)} = [a_{nm,1}^{(l)}, \ldots, a_{nm,L}^{(l)}]^T$ written by

$$\mathcal{N}(\mathbf{a}_{nm}^{(l)}|\boldsymbol{\mu}_{a_{nm}}^{(l-1)} = 0, (\mathbf{R}_{a_{nm}}^{(l-1)})^{-1} = \mathbf{K}_{a_{nm}}^{(l-1)}) \tag{19}$$

which is a Gaussian with zero mean and an $L \times L$ covariance matrix $\mathbf{K}_{a_{nm}}^{(l-1)}$ with the (t, τ)th entry

$$[\mathbf{K}_{a_{nm}}^{(l-1)}]_{t\tau} = \kappa(\vec{\mathbf{a}}_{nm,t-1}^{(l)}, \vec{\mathbf{a}}_{nm,\tau-1}^{(l)}) + \delta_{t\tau}. \tag{20}$$

Regarding the source signals, we could similarly incorporate a GP prior to represent the temporal structure of time-varying source samples $\{s_{m,t}^{(l)}\}$ within a frame. A latent function $f(\vec{\mathbf{s}}_{m,t-1}^{(l)})$ of previous p samples $\vec{\mathbf{s}}_{m,t-1}^{(l)} = [s_{m,t-1}^{(l)}, \ldots, s_{m,t-p}^{(l)}]^T$ is employed to predict the current source sample $s_{m,t}^{(l)}$ and is distributed by a zero-mean Gaussian prior $\mathcal{N}(f(\vec{\mathbf{s}}_{m,t-1}^{(l)}))|0, \kappa(\vec{\mathbf{s}}_{m,t-1}^{(l)}, \vec{\mathbf{s}}_{m,\tau-1}^{(l)}))$ where $\kappa(\vec{\mathbf{s}}_{m,t-1}^{(l)}, \vec{\mathbf{s}}_{\tau,t-1}^{(l)})$ denotes the variance calculated by the exponential-quadratic kernel function given in (18) but substituting the corresponding kernel parameters $\{\lambda_{s_m}^{(l-1)}, \xi_{s_m}^{(l-1)}\}$. The GP prior density $p(\mathbf{s}_m^{(l)}|\boldsymbol{\mu}_{s_m}^{(l-1)}, \mathbf{R}_{s_m}^{(l-1)})$ of L source samples $\mathbf{s}_m^{(l)} = [s_{m,1}^{(l)}, \ldots, s_{m,L}^{(l)}]^T$ is similarly obtained by $\mathcal{N}(\mathbf{s}_m^{(l)}|\boldsymbol{\mu}_{s_m}^{(l-1)} = 0, (\mathbf{R}_{s_m}^{(l-1)})^{-1} = \mathbf{K}_{s_m}^{(l-1)})$ with the (t, τ)th entry of covariance matrix given by $[\mathbf{K}_{s_m}^{(l-1)}]_{t\tau} = \kappa(\vec{\mathbf{s}}_{m,t-1}^{(l)}, \vec{\mathbf{s}}_{m,\tau-1}^{(l)}) + \delta_{t\tau}$.

Fig. 4 displays the graphical representation of the OLGP-ICA model. The noisy ICA model is used with a noise vector $\boldsymbol{\varepsilon}_t^{(l)}$, which is assumed to be Gaussian distributed $\mathcal{N}(\boldsymbol{\varepsilon}_t^{(l)}|0, (\mathbf{B}^{(l)})^{-1})$ with zero mean and diagonal precision matrix $\mathbf{B}^{(l)} = \text{diag}\{\beta_n^{(l)}\}$. The precision parameter $\beta_n^{(l)}$ is generated by a gamma prior given by $\text{gamma}(\beta_n^{(l)}|u_{\beta_n}^{(l-1)}, \omega_{\beta_n}^{(l-1)})$. Hence, the OLGP-ICA model parameters are formed by $\mathbf{\Theta}^{(l)} = \{\mathbf{A}^{(l)}, \mathbf{E}^{(l)}, \mathbf{B}^{(l)}, \mathbf{S}^{(l)}\}$ and their hyperparameters $\mathbf{\Phi}^{(l)} = \{\mathbf{M}_a^{(l)}, \mathbf{R}_a^{(l)}, \mathbf{\Lambda}_a^{(l)}, \mathbf{\Xi}_a^{(l)}, \mathbf{u}^{(l)}, \boldsymbol{\omega}^{(l)}, \mathbf{M}_s^{(l)}, \mathbf{R}_s^{(l)}, \boldsymbol{\lambda}_s^{(l)}, \boldsymbol{\xi}_s^{(l)}\}$ consist of Gaussian parameters of mixing coefficients $\{\mathbf{M}_a^{(l)} = \{\boldsymbol{\mu}_{a_{nm}}^{(l)}\}, \mathbf{R}_a^{(l)} = \{\mathbf{R}_{a_{nm}}^{(l)}\}\}$ and source signals $\{\mathbf{M}_s^{(l)} = \{\boldsymbol{\mu}_{s_m}^{(l)}\}, \mathbf{R}_s^{(l)} = \{\mathbf{R}_{s_m}^{(l)}\}\}$, gamma parameters of noise signals $\{\mathbf{u}^{(l)} = \{u_{\beta_n}^{(l)}\}, \boldsymbol{\omega}^{(l)} = \{\omega_{\beta_n}^{(l)}\}\}$, and kernel parameters of mixing coefficients $\{\mathbf{\Lambda}_a^{(l)} = \{\lambda_{a_{nm}}^{(l)}\}, \mathbf{\Xi}_a^{(l)} = \{\xi_{a_{nm}}^{(l)}\}\}$ and source signals $\{\boldsymbol{\lambda}_s^{(l)} = \{\lambda_{s_m}^{(l)}\}, \boldsymbol{\xi}_s^{(l)} = \{\xi_{s_m}^{(l)}\}\}$. This paper highlights the nonstationary and temporally correlated source separation by using online GP. The proposed OLGP-ICA differs from the GP factor analysis (GP-FA) [32], which was developed

TABLE I
SEQUENTIAL AND VARIATIONAL BAYESIAN LEARNING ALGORITHM

Initialize With $\mathbf{\Phi}^{(0)}$ and $\mathbf{\Theta}^{(0)}$
For Each Data Segment $\mathbf{X}^{(l)} = \{\mathbf{x}_t^{(l)}\}$
for each VB-EM iteration
VB-E step: update variational distributions
for each time sample $\mathbf{x}_t^{(l)}$, $t = 1, \ldots, L$
accumulate sufficient statistics
find $q(\mathbf{\Theta}^{(l)}\|\mathbf{\Phi}^{(l)})$ and update $\mathbf{\Phi}^{(l)} \leftarrow \mathbf{\Phi}^{(l-1)}$
VB-M step: estimate model parameters
find distribution mode and update $\mathbf{\Theta}^{(l+1)} \leftarrow \mathbf{\Theta}^{(l)}$
check if $\|\mathbf{\Theta}^{(l+1)} - \mathbf{\Theta}^{(l)}\|$ is small enough
$l \leftarrow l + 1$

for spatiotemporal data modeling. The factor loading matrix and the common factors in GP-FA were time invariant. Notably, the scaling hyperparameter $\lambda_{a_{nm}}^{(l)}$ in OLGP-ICA considerably affects the representation of temporal structure of $\mathbf{a}_{nm}^{(l)}$. If the estimated $\lambda_{a_{nm}}^{(l)}$ is very small, it implies that the associated $\vec{\mathbf{a}}_{nm,t-1}^{(l)}$ have little effect on the prediction of $a_{nm,t}^{(l)}$. Estimating $\lambda_{a_{nm}}^{(l)}$ is equivalent to performing the ARD scheme [5] or reflecting the activities of source signals. When the source signal is abruptly inactive, the corresponding mixing coefficients turn out to be zero. The previous mixing coefficients have little impact on the prediction of current mixing coefficient. The status of source signals at different frames is determined according to the updated scaling parameter $\lambda_{a_{nm}}^{(l)}$.

IV. SEQUENTIAL AND VARIATIONAL ICA ALGORITHMS

We present an online Bayesian learning procedure for NB-ICA and OLGP-ICA algorithms by maximizing likelihood function $p(\mathbf{X}^{(l)}|\mathbf{\Phi}^{(l-1)})$, which is marginalized over latent variables or model parameters $\mathbf{\Theta}^{(l)}$. As summarized in Table I, at each sequential learning epoch l, we perform the variational Bayesian expectation-maximization (VB-EM) iteration by using data segment $\mathbf{X}^{(l)}$. For each VB-EM iteration, the variational distributions of individual parameters are updated to find new hyperparameters or variational parameters $\mathbf{\Phi}^{(l)}$ in VB-E step. New model parameters $\mathbf{\Theta}^{(l+1)}$ are then estimated in the VB-M step. The learning procedure is detailed in the following.

A. Sequential VB Inference for NB-ICA

First of all, the sequential VB inference procedure is developed for NB-ICA with model parameters $\mathbf{\Theta}^{(l)} = \{\mathbf{A}^{(l)}, \boldsymbol{\alpha}^{(l)}\mathbf{H}^{(l)}, \mathbf{E}^{(l)}, \mathbf{B}^{(l)}, \mathbf{S}^{(l)}, \mathbf{\Pi}^{(l)}, \mathbf{M}^{(l)}, \mathbf{R}^{(l)}\}$ and hyperparameters $\mathbf{\Phi}^{(l)} = \{\boldsymbol{\rho}^{(l)}, \mathbf{V}^{(l)}, \mathbf{u}^{(l)}, \boldsymbol{\omega}^{(l)}, \mathbf{\Phi}_{\pi}^{(l)}, \mathbf{\Phi}_{m}^{(l)}, \mathbf{\Phi}_{r}^{(l)}\}$. NB-ICA parameters $\mathbf{\Theta}^{(l)}$ are all latent variables. Since the exact inference using posterior distribution $p(\mathbf{\Theta}^{(l)}|\mathbf{X}^{(l)}, \mathbf{\Phi}^{(l-1)})$ is intractable due to coupling among latent variables, the VB-EM procedure [8], [31], [52] is applied to conduct approximate inference by maximizing the negative free energy or the lower bound of the logarithm of marginal likelihood. The marginal likelihood $p(\mathbf{X}^{(l)}|\mathbf{\Phi}^{(l-1)})$ is calculated by taking multiple integrals over different latent variables in $\mathbf{\Theta}^{(l)}$ by using L data samples $\mathbf{X}^{(l)} = \{\mathbf{x}_t^{(l)}\}$

$$\begin{aligned}\prod_{t=1}^{L}\int & p(\mathbf{x}_t^{(l)}|\mathbf{A}^{(l)}, \mathbf{s}_t^{(l)}, \boldsymbol{\varepsilon}_t^{(l)}) p(\mathbf{A}^{(l)}|\boldsymbol{\alpha}^{(l)}, \mathbf{H}^{(l)}) \\ & \times p(\boldsymbol{\alpha}^{(l)}\mathbf{H}^{(l)}|\boldsymbol{\rho}^{(l-1)}, \mathbf{V}^{(l-1)}) p(\boldsymbol{\varepsilon}_t^{(l)}|\mathbf{B}^{(l)}) \\ & \times p(\mathbf{B}^{(l)}|\mathbf{u}^{(l-1)}, \boldsymbol{\omega}^{(l-1)}) p(\mathbf{s}_t^{(l)}|\mathbf{\Pi}^{(l)}, \mathbf{M}^{(l)}, \mathbf{R}^{(l)}) \\ & \times p(\mathbf{\Pi}^{(l)}|\mathbf{\Phi}_{\pi}^{(l-1)}) p(\mathbf{M}^{(l)}|\mathbf{\Phi}_{m}^{(l-1)}) p(\mathbf{R}^{(l)}|\mathbf{\Phi}_{r}^{(l-1)}) \\ & d\mathbf{A}^{(l)} d(\boldsymbol{\alpha}^{(l)}\mathbf{H}^{(l)}) d\boldsymbol{\varepsilon}_t^{(l)} d\mathbf{B}^{(l)} d\mathbf{s}_t^{(l)} d\mathbf{\Pi}^{(l)} d\mathbf{M}^{(l)} d\mathbf{R}^{(l)}. \end{aligned} \quad (21)$$

The updated hyperparameters $\mathbf{\Phi}^{(l-1)}$ from previous frames $\mathcal{X}^{(l-1)}$ are given in VB-EM procedure. A variational distribution $q(\mathbf{\Theta}^{(l)})$ is used to approximate the true posterior distribution $p(\mathbf{\Theta}^{(l)}|\mathbf{X}^{(l)}, \mathbf{\Phi}^{(l-1)})$ at each frame. Maximizing the lower bound of $\log p(\mathbf{X}^{(l)}|\mathbf{\Phi}^{(l-1)})$ is equivalent to maximizing the expectation of log likelihood $E_q[\log p(\mathbf{X}^{(l)}|\mathbf{\Theta}^{(l)}, \mathbf{\Phi}^{(l-1)})]$ over $q(\mathbf{\Theta}^{(l)})$ or minimizing the Kullback–Leibler divergence between $q(\mathbf{\Theta}^{(l)})$ and $p(\mathbf{\Theta}^{(l)}|\mathbf{X}^{(l)}, \mathbf{\Phi}^{(l-1)})$. Considering the factorized variational inference using

$$\begin{aligned} q(\mathbf{\Theta}^{(l)}) = \prod_j q(\mathbf{\Theta}_j^{(l)}) = & q(\mathbf{A}^{(l)}) q(\alpha^{(l)}\mathbf{H}^{(l)}) q(\mathbf{E}^{(l)}) \\ & \times q(\mathbf{B}^{(l)}) q(\mathbf{S}^{(l)}) q(\mathbf{\Pi}^{(l)}) q(\mathbf{M}^{(l)}) q(\mathbf{R}^{(l)}) \end{aligned} \quad (22)$$

the lower bound of $\log p(\mathbf{X}^{(l)}|\mathbf{\Phi}^{(l-1)})$ is expanded as

$$\begin{aligned} & E_q[\log p(\mathbf{X}^{(l)}|\mathbf{A}^{(l)}, \mathbf{S}^{(l)}, \mathbf{E}^{(l)})] \\ & \quad + E_q[\log p(\mathbf{A}^{(l)}|\boldsymbol{\alpha}^{(l)}, \mathbf{H}^{(l)})] + S(q(\mathbf{A}^{(l)})) \\ & \quad + E_q[\log p(\boldsymbol{\alpha}^{(l)}\mathbf{H}^{(l)}|\boldsymbol{\rho}^{(l-1)}, \mathbf{V}^{(l-1)})] + S(q(\boldsymbol{\alpha}^{(l)}\mathbf{H}^{(l)})) \\ & \quad + E_q[\log p(\mathbf{E}^{(l)}|\mathbf{B}^{(l)})] + S(q(\mathbf{E}^{(l)})) \\ & \quad + E_q[\log p(\mathbf{B}^{(l)}|\mathbf{u}^{(l-1)}, \boldsymbol{\omega}^{(l-1)})] + S(q(\mathbf{B}^{(l)})) \\ & \quad + E_q[\log p(\mathbf{S}^{(l)}|\mathbf{\Pi}^{(l)}, \mathbf{M}^{(l)}, \mathbf{R}^{(l)})] + S(q(\mathbf{S}^{(l)})) \\ & \quad + E_q[\log p(\mathbf{\Pi}^{(l)}|\mathbf{\Phi}_{\pi}^{(l-1)})] + S(q(\mathbf{\Pi}^{(l)})) \\ & \quad + E_q[\log p(\mathbf{M}^{(l)}|\mathbf{\Phi}_{m}^{(l-1)})] + S(q(\mathbf{M}^{(l)})) \\ & \quad + E_q[\log p(\mathbf{R}^{(l)}|\mathbf{\Phi}_{r}^{(l-1)})] + S(q(\mathbf{R}^{(l)})) \end{aligned} \quad (23)$$

where $S(q(\cdot))$ denotes the entropy of a distribution $q(\cdot)$. By taking partial differential of (23) with respect to the jth variational distribution $q(\mathbf{\Theta}_j^{(l)})$ and setting it to zero, the optimal variational distribution is derived and expressed in a general form [5], [31]

$$\log \hat{q}(\mathbf{\Theta}_j^{(l)}) \propto E_{q(\Theta \neq \Theta_j)}[\log p(\mathbf{X}^{(l)}, \mathbf{\Theta}^{(l)}|\mathbf{\Phi}^{(l-1)})] \quad (24)$$

where the expectation is taken with respect to all of the other factors of previous estimates $q(\mathbf{\Theta}^{(l)} \neq \mathbf{\Theta}_j^{(l)})$. Finding the variational distributions, $\hat{q}(\mathbf{\Theta}^{(l)})$ is equivalent to updating the hyperparameters $\mathbf{\Phi}^{(l-1)} \rightarrow \mathbf{\Phi}^{(l)}$ in VB-E step. The solutions to hyperparameters $\{\mathbf{u}^{(l)}, \boldsymbol{\omega}^{(l)}, \mathbf{\Phi}_{\pi}^{(l)}, \mathbf{\Phi}_{m}^{(l)}, \mathbf{\Phi}_{r}^{(l)}\}$ have been derived in [16]. Here, the solution to hyperparameters $\{\boldsymbol{\rho}^{(l)}, \boldsymbol{V}^{(l)}\}$ is obtained according to (24) by applying the prior densities of (14) and (15). The optimal variational distribution is derived

by

$$\begin{aligned}\log \prod_{m=1}^{M} \hat{q}(\alpha_m^{(l)}\mathbf{H}_m^{(l)}) &\propto E_{q(\mathbf{A}^{(l)})}[\log p(\mathbf{A}^{(l)}|\boldsymbol{\alpha}^{(l)}, \mathbf{H}^{(l)})] \\ &+ \log p(\boldsymbol{\alpha}^{(l)}\mathbf{H}^{(l)}|\boldsymbol{\rho}^{(l-1)}, \mathbf{V}^{(l-1)}) \\ &= \sum_{m=1}^{M}\{E_{q(\mathbf{a}_m^{(l)})}[\log p(\mathbf{a}_m^{(l)}|\alpha_m^{(l)}, \mathbf{H}_m^{(l)})] \\ &+ \log p(\alpha_m^{(l)}\mathbf{H}_m^{(l)}|\rho_m^{(l-1)}, \mathbf{V}_m^{(l-1)})\} \\ &\propto \sum_{m=1}^{M}\left\{\begin{array}{c}\frac{\rho_m^{(l-1)}-N}{2}\log|\alpha_m^{(l)}\mathbf{H}_m^{(l)}| \\ -\frac{1}{2}\left[\mathrm{Tr}[(E_{q(\mathbf{a}_m^{(l)})}[\mathbf{a}_m^{(l)}(\mathbf{a}_m^{(l)})^T]+(\mathbf{V}_m^{(l-1)})^{-1})\alpha_m^{(l)}\mathbf{H}_m^{(l)}]\right]\end{array}\right\}. \end{aligned} \tag{25}$$

The variational distribution $\hat{q}(\alpha_m^{(l)}\mathbf{H}_m^{(l)}|\rho_m^{(l)}, V_m^{(l)})$ is seen as a new Wishart distribution with the updated hyperparameters

$$\rho_m^{(l)} = \rho_m^{(l-1)} + 1 \tag{26}$$

$$\mathbf{V}_m^{(l)} = (E_{q(\mathbf{a}_m^{(l)})}[\mathbf{a}_m^{(l)}(\mathbf{a}_m^{(l)})^T] + (\mathbf{V}_m^{(l-1)})^{-1})^{-1} \tag{27}$$

which are used for online learning of individual mixing vector $\mathbf{a}_m^{(l-1)} \rightarrow \mathbf{a}_m^{(l)}$. The NB-ICA procedure is implemented by continuously applying current frame $\mathbf{X}^{(l)}$ to refresh the hyperparameters $\boldsymbol{\Phi}^{(l-1)} = \{\rho_m^{(l-1)}, \mathbf{V}_m^{(l-1)}\} \rightarrow \boldsymbol{\Phi}^{(l)} = \{\rho_m^{(l)}, \mathbf{V}_m^{(l)}\}$, which are used at next learning epoch when new frame $\mathbf{X}^{(l+1)}$ is enrolled. In the VB-M step, the model parameters are accordingly updated via $\boldsymbol{\Theta}^{(l)} \rightarrow \boldsymbol{\Theta}^{(l+1)}$ by using the updated modes of variational distributions, namely, $\boldsymbol{\Phi}^{(l)}$.

B. Sequential VB Inference for OLGP-ICA

In NB-ICA approach, the mixing matrix and source signals are generated by Gaussian distribution and MoG distributions, respectively. The OLGP-ICA algorithm is developed by exploring the time structures of mixing matrix $\mathbf{A}_t^{(l)}$ and source signals $\mathbf{s}_t^{(l)}$ through the GP, which is a generalization of Gaussian distribution for time-varying random variables. The latent functions of $\{\mathbf{A}_t^{(l)}, \mathbf{s}_t^{(l)}\}$ are expressed by Gaussian priors using the kernel parameters. Based on OLGP-ICA approach, the system parameters $\boldsymbol{\Theta}^{(l)} = \{\mathbf{A}^{(l)}, \mathbf{E}^{(l)}, \mathbf{B}^{(l)}, \mathbf{S}^{(l)}\}$ and their hyperparameters $\boldsymbol{\Phi}^{(l)} = \{\mathbf{M}_a^{(l)}, \mathbf{R}_a^{(l)}, \boldsymbol{\Lambda}_a^{(l)}, \boldsymbol{\Xi}_a^{(l)}, \mathbf{u}^{(l)}, \boldsymbol{\omega}^{(l)}, \mathbf{M}_s^{(l)}, \mathbf{R}_s^{(l)}, \boldsymbol{\lambda}_s^{(l)}, \boldsymbol{\xi}_s^{(l)}\}$ at each frame are inferred by a VB-EM procedure. Again, the lower bound of marginal likelihood $p(\mathbf{X}^{(l)}|\boldsymbol{\Phi}^{(l-1)})$ is maximized to find the optimal variational distribution $\hat{q}(\boldsymbol{\Theta}^{(l)})$ with the updated hyperparameters $\boldsymbol{\Phi}^{(l-1)} \rightarrow \boldsymbol{\Phi}^{(l)}$. VB-E step is performed. The solution to hyperparameters $\{\mathbf{u}^{(l)}, \boldsymbol{\omega}^{(l)}\}$ for noise signals has been derived in [16]. In this paper, we formulate the variational distributions of mixing coefficients $\mathbf{a}_{nm}^{(l)}$ and source signals $\mathbf{s}_m^{(l)}$ at frame l based on the Gaussian process. According to the general solution of (24), the optimal variational distribution of mixing coefficients is yielded by

$$\begin{aligned}\log \hat{q}(\mathbf{a}_{nm}^{(l)}) \propto\ & E_{q(\Theta\neq A)}[\log p(\mathbf{X}^{(l)}|\mathbf{a}_{nm}^{(l)}, \mathbf{s}_m^{(l)}, \boldsymbol{\varepsilon}_n^{(l)})] \\ &+ \log p(\mathbf{a}_{nm}^{(l)}|\boldsymbol{\mu}_{a_{nm}}^{(l-1)}, \mathbf{R}_{a_{nm}}^{(l-1)})\end{aligned} \tag{28}$$

where $\boldsymbol{\varepsilon}_n^{(l)} = [\boldsymbol{\varepsilon}_{n,1}^{(l)}, \ldots, \boldsymbol{\varepsilon}_{n,L}^{(l)}]$. In (28), the first term is an expectation function operated over all variational distributions $q(\boldsymbol{\Theta}^{(l)} \neq \mathbf{A}_t^{(l)})$ except that of $\mathbf{A}_t^{(l)}$ and the second term is a Gaussian prior given by (19). The first term should be manipulated as a quadratic function of $\mathbf{a}_{nm}^{(l)}$ so that two terms in right-hand side (RHS) of (28) can be combined into an exponent of a new Gaussian with the updated hyperparameters $\{\boldsymbol{\mu}_{a_{nm}}^{(l-1)}, \mathbf{R}_{a_{nm}}^{(l-1)}\} \rightarrow \{\boldsymbol{\mu}_{a_{nm}}^{(l)}, \mathbf{R}_{a_{nm}}^{(l)}\}$. To do so, a new expression of Gaussian distribution is arranged as [32]

$$\begin{aligned}&\exp\{E_{q(\Theta\neq A)}[\log p(\mathbf{X}^{(l)}|\mathbf{a}_{nm}^{(l)}, \mathbf{s}_m^{(l)}, \boldsymbol{\varepsilon}_n^{(l)})]\} \\ &\quad\propto \mathcal{N}((\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1}\tilde{\mathbf{x}}_{a_{nm}}^{(l)}|\mathbf{a}_{nm}^{(l)}, (\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1})\end{aligned} \tag{29}$$

where $\tilde{\mathbf{x}}_{a_{nm}}^{(l)}$ denotes an L dimensional vector with the tth entry

$$\begin{aligned}\tilde{x}_{a_{nm},t}^{(l)} =\ & E_{q(\Theta\neq A)}\left[\beta_{n,t}^{(l)}\right] E_{q(\Theta\neq A)}\left[s_{m,t}^{(l)}\right] \\ &\times\left(x_{n,t}^{(l)} - \sum_{k\neq m}^{M} E_{q(\Theta\neq A)}\left[a_{nk,t}^{(l)}\right] E_{q(\Theta\neq A)}\left[s_{k,t}^{(l)}\right]\right)\end{aligned} \tag{30}$$

and $\boldsymbol{\Psi}_{a_{nm}}^{(l)}$ denotes an $L \times L$ diagonal matrix with tth diagonal entry

$$[\boldsymbol{\Psi}_{a_{nm}}^{(l)}]_{tt} = E_{q(\Theta\neq A)}[\beta_{n,t}^{(l)}]E_{q(\Theta\neq A)}[(s_{m,t}^{(l)})^2]. \tag{31}$$

Equation (29) is a Gaussian likelihood function of the transformed observation vector $(\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1}\tilde{\mathbf{x}}_{a_{nm}}^{(l)}$, which is expressed as a quadratic function of $\mathbf{a}_{nm}^{(l)}$. By substituting the likelihood function of (29) and the prior density of (19) into (28), the two exponents of quadratic functions of $\mathbf{a}_{nm}^{(l)}$ are summed up to achieve the optimal variational distribution $\hat{q}(\mathbf{a}_{nm}^{(l)}|\boldsymbol{\mu}_{a_{nm}}^{(l)}, \mathbf{R}_{a_{nm}}^{(l)})$, which acts as a posterior distribution and is expressed as a new Gaussian distribution with the updated hyperparameters

$$\boldsymbol{\mu}_{a_{nm}}^{(l)} = (\mathbf{R}_{a_{nm}}^{(l-1)})^{-1}(\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1}\tilde{\mathbf{x}}_{a_{nm}}^{(l)} \tag{32}$$

$$\mathbf{R}_{a_{nm}}^{(l)} = (\boldsymbol{\Psi}_{a_{nm}}^{(l)}(\mathbf{R}_{a_{nm}}^{(l-1)})^{-1})^{-1}. \tag{33}$$

Following the perspective of online Bayesian learning, the OLGP-ICA procedure is established by combining a Gaussian likelihood and a GP prior and reproducing a variational posterior $\hat{q}(\mathbf{a}_{nm}^{(l)}|\boldsymbol{\mu}_{a_{nm}}^{(l)}, \mathbf{R}_{a_{nm}}^{(l)})$, which is also Gaussian. The evolution of hyperparameters is performed by $\{\mathbf{M}_a^{(l-1)}, \mathbf{R}_a^{(l-1)}\} \rightarrow \{\mathbf{M}_a^{(l)}, \mathbf{R}_a^{(l)}\}$. The Appendix addresses the updating formulas for the remaining OLGP-ICA hyperparameters $\{\mathbf{M}_s^{(l)}, \mathbf{R}_s^{(l)}\}$ and $\{\boldsymbol{\Lambda}_a^{(l)}, \boldsymbol{\Xi}_a^{(l)}, \boldsymbol{\lambda}_s^{(l)}, \boldsymbol{\xi}_s^{(l)}\}$. Finally, an online learning algorithm is implemented by continuously applying current frame $\mathbf{X}^{(l)}$ to update the hyperparameters $\boldsymbol{\Phi}^{(l-1)} \rightarrow \boldsymbol{\Phi}^{(l)} = \{\mathbf{M}_a^{(l)}, \mathbf{R}_a^{(l)}, \boldsymbol{\Lambda}_a^{(l)}, \boldsymbol{\Xi}_a^{(l)}, \mathbf{u}^{(l)}, \boldsymbol{\omega}^{(l)}, \mathbf{M}_s^{(l)}, \mathbf{R}_s^{(l)}, \boldsymbol{\lambda}_s^{(l)}, \boldsymbol{\xi}_s^{(l)}\}$ and propagate them to next learning epoch by applying $\mathbf{X}^{(l+1)}$. When implementing the OLGP-ICA procedure, we start from the initial hyperparameters $\boldsymbol{\Phi}^{(0)}$ and iteratively update them in turn and replace the hyperparameters by new estimates $\boldsymbol{\Phi}^{(l-1)} \rightarrow \boldsymbol{\Phi}^{(l)}$. At each re-estimation iteration, the lower bound is increased until the variational posterior reaches its maximum. New hyperparameters $\boldsymbol{\Phi}^{(l)}$ at frame l are propagated to the next frame $l+1$ and act as new statistics of the priors for sequential and VB learning. Again, given the new hyperparameters, the model parameters of OLGP-ICA are estimated in VB-M step.

C. Sequential Monte Carlo ICA

In addition to the deterministic inference using VB, the stochastic inference based on MC method and importance sampling was developed to establish ICA model for image separation [44]. Typically, VB and MC inferences are algorithmically similar but theoretically different [7]. The greedy algorithm, such as VB and the local algorithm, such as MC share similar local updating strategy for Bayesian learning and can get trapped in a local mode of the posterior, depending upon the starting configuration. In general, VB converges quickly to a nearby mode while MC is advantageous with good generality and robustness. In [7], VB performed slightly better than MC in terms of different performance criteria. The prior of source signal using Student's t distribution was introduced. In [37], MC inference was developed to construct the nonnegative matrix factorization algorithm for nonnegative source separation. All of these MC methods were designed for batch learning. This paper investigates the online learning strategy. In [17], the sequential importance sampling was employed in the jump Markov linear systems (JMLS) where the time-varying state parameters were recursively computed by particle filters. JMLS were exploited to build generic systems in presence of continuous-state process as well as discrete-state process [17]. Nevertheless, ICA system only involved the continuous-state process. The sequential MC-based ICA (SMC-ICA) algorithm was accordingly developed [1], [6]. In [22], the particle filtering was performed for non-Gaussian AR process but without dealing with ICA problem. In [41], the MC method was employed for Bayesian learning by using conjugate priors while the temporally correlated source separation was not considered.

Using SMC-ICA, the ICA model in (8) is rewritten in terms of a continuous-state equation and an observation equation

$$\boldsymbol{a}^{(l)} = \boldsymbol{a}^{(l-1)} + \mathbf{v}^{(l)} \tag{34}$$

$$\mathbf{x}_t^{(l)} = \mathbf{C}_t^{(l)} \boldsymbol{a}^{(l)} + \boldsymbol{\varepsilon}_t^{(l)} \tag{35}$$

where $\boldsymbol{a}^{(l)} = \text{vec}\{\mathbf{A}^{(l)}\}$ denotes an $NM \times 1$ vector consisting of mixing coefficients so that $[\boldsymbol{a}^{(l)}]_{N(m-1)+n} = a_{nm}^{(l)}$, $\mathbf{v}^{(l)}$ denotes a vector of zero-mean Gaussian noise and $\mathbf{C}_t^{(l)} = (\mathbf{s}_t^{(l)})^T \otimes \mathbf{I}_N$ denotes an $N \times NM$ matrix of source signals. The source signals are modeled by a mixture of Gaussians as given in (11) where the transition probability from mixture variables $z_{m,t}^{(l-1)} = j$ to $z_{m,t}^{(l)} = k$ with $1 \le j, k \le K$ are additionally defined by $p(z_{m,t}^{(l)} = k | z_{m,t}^{(l-1)} = j) = \tau_{mjk}^{(l)}$. This probability is used to indicate which Gaussian component is active at time t for source signal m. We construct the SMC-ICA parameters as $\boldsymbol{\Theta}^{(l)} = \{\{\mathbf{s}_t^{(l)}\}, \{z_{m,t}^{(l)}\}, \{\mu_{mk}^{(l)}\}, \{\gamma_{mk}^{(l)}\}, \{\tau_{mjk}^{(l)}\}\}$. In this particle filter, the particles contain $\{\boldsymbol{a}_q^{(l)}, \boldsymbol{\Theta}_q^{(l)}, 1 \le q \le Q\}$, which are sampled from $\mathbf{X}^{(1:l)} = \mathcal{X}^{(l)} = \{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(l)}\}$ according to the posterior density

$$p(\boldsymbol{a}^{(l)}, \Theta^{(0:l)} | \mathbf{X}^{(1:l)}) = p(\boldsymbol{a}^{(l)} | \Theta^{(0:l)}, \mathbf{X}^{(1:l)}) p(\Theta^{(0:l)} | \mathbf{X}^{(1:l)}). \tag{36}$$

Given an approximation of $p(\boldsymbol{\Theta}^{(0:l)} | \mathbf{X}^{(1:l)})$, the mixing coefficients are determined by $p(\boldsymbol{a}^{(l)} | \boldsymbol{\Theta}^{(0:l)}, \mathbf{X}^{(1:l)})$, which is a Gaussian distribution and can be recursively estimated in closed form by using the Kalman filter based on (34), (35). The particle filtering is then performed according to the posterior $p(\boldsymbol{\Theta}^{(0:l)} | \mathbf{X}^{(1:l)})$, which is recursively calculated via a sub-optimal method using the importance distribution [1], [6]

$$\pi(\Theta^{(0:l)} | \mathbf{X}^{(1:l)}) = \pi(\Theta^{(0)}) \prod_{j=1}^{l} \pi(\Theta^{(j)} | \Theta^{(0:j-1)}, \mathbf{X}^{(1:j)}). \tag{37}$$

The prior importance function $\pi(\boldsymbol{\Theta}^{(l)} | \boldsymbol{\Theta}^{(l-1)})$ is expressed by

$$\begin{aligned} p(\{\mathbf{s}_t^{(l)}\} | \{z_{m,t}^{(l)}\}, \{\mu_{mk}^{(l)}\}, \{\gamma_{mk}^{(l)}\}) p(\{\mu_{mk}^{(l)}\} | \{\mu_{mk}^{(l-1)}\}) \\ \times p(\{\gamma_{mk}^{(l)}\} | \{\gamma_{mk}^{(l-1)}\}) p(\{z_{m,t}^{(l)}\} | \{z_{m,t}^{(l-1)}\}, \{\tau_{mjk}^{(l)}\}) \\ \times p(\{\tau_{mjk}^{(l)}\} | \{\tau_{mjk}^{(l-1)}\}) \end{aligned} \tag{38}$$

where $\{\mu_{mk}^{(l)}\}$ and $\{\log \gamma_{mk}^{(l)}\}$ are drawn at each frame l from Gaussian distributions with the means at previous value of the respective particles at frame $l-1$ and the variances as determined in [1]and [6]. The particle filtering algorithm is implemented for ICA using the following two steps [1], [6], and [17].

Sequential Importance Sampling Step

1) For $q = 1, \ldots, Q$, sample $\tilde{\Theta}_q^{(l)}$ from the distribution $\pi(\Theta^{(l)} | \Theta_q^{(0:l-1)}, \mathbf{X}^{(1:l)})$ and set $\tilde{\Theta}_q^{(0:l)} = \{\Theta_q^{(0:l-1)}, \tilde{\Theta}_q^{(l)}\}$.
2) For $q = 1, \ldots, Q$, evaluate the importance weights up to a normalized constant

$$w_q^{(l)} \propto \frac{p(\mathbf{X}^{(l)} | \tilde{\Theta}_q^{(0:l)}, \mathbf{X}^{(1:l-1)}) p(\tilde{\Theta}_q^{(l)} | \tilde{\Theta}_q^{(l-1)})}{\pi(\tilde{\Theta}_q^{(l)} | \tilde{\Theta}_q^{(0:l-1)}, \mathbf{X}^{(1:l)})}. \tag{39}$$

3) For $q = 1, \ldots, Q$, normalize the importance weights by

$$\tilde{w}_q^{(l)} \propto \left[\sum_{j=1}^{Q} w_j^{(l)} \right]^{-1} w_q^{(l)}. \tag{40}$$

Selection Step

1) Discard/multiply particles with low/high normalized importance weights to obtain particles $\{\Theta_q^{(0:l)}, q = 1, \ldots, Q\}$.

Given these particles and normalized importance weights, the SMC-ICA parameters $\Theta^{(l)}$ or the source signals $\{\mathbf{s}_t^{(l)}\}$ are estimated by calculating the expectations

$$E_{p_Q(\Theta^{(0:l)} | \mathbf{X}^{(1:l)})}[f^{(l)}(\Theta^{(0:l)})] = \sum_{q=1}^{Q} f^{(l)}(\Theta_q^{(l)}) \tilde{w}_q^{(l)}. \tag{41}$$

If $f^{(l)}$ is an identify function, (41) is used to obtain the minimum mean square estimate of $\Theta^{(l)}$. Notably, SMC-ICA incrementally compensates the variations of mixing coefficients as well as source signals. In the implementation, we refer to the experimental conditions as given in [1] and [6].

V. Experiments

A. Experimental Setup

In the experiments, the proposed NB-ICA and OLGP-ICA algorithms are evaluated for nonstationary source separation by using real-world audio signals. The speech and music signals were sampled from the ICA'99 Test Sets, which are available at: http://sound.media.mit.edu/ica-bench/. Different scenarios are simulated by using dynamic mixing coefficients and dynamic source signals. Fig. 5(a) shows an example of two waveforms of dynamic source signals with 5 s containing two different speakers and one music source. In the first channel, a male speaker was speaking and inactive at 1.5 s and then replaced by a music source at 2.5 s. In this scenario, the distribution of source signal was changed. The switching and moving sources were simulated. In the second channel, a different male speaker was speaking and inactive at 2.5 s and then speaking again at 3.5 s. The presence and absence of the same speaker was simulated. In addition, the nonstationary environments were simulated by using a time-varying mixing matrix $\mathbf{A}_t = [[\cos(2\pi f_1 t) \ -\sin(2\pi f_1 t)]^T [\sin(2\pi f_2 t) \ \cos(2\pi f_2 t)]^T]$ where f_1 and f_2 denote the changing rate of mixing coefficients. The entries of the mixing matrix revealed the spatial information about the relation between sources and sensors. The first and the second columns of $\mathbf{A}_t$ reflected the relation of the first and the second source signals to the sensors, respectively. The changing rates $f_1 = 1/20$ Hz and $f_2 = 1/10$ Hz were used.

We conducted a comparison between the proposed NB-ICA and OLGP-ICA and the other seven ICA methods, including VB-ICA [31], BICA-HMM [15], S-ICA [23], GP-ICA [39], online VB-ICA [24], NS-ICA [18] and SMC-ICA [1], [6]. The VB-ICA [31] performed the batch learning and did not deal with the nonstationary signals. The Bayesian ICA with hidden Markov sources (BICA-HMM) [15] was a batch learning method by applying an HMM to represent the switching sources. The S-ICA [23] was a batch learning method, tackling the scenario of abrupt active and inactive sources. The scenario of moving sources was not considered in [15], [23]. The GP-ICA [39] explored the temporal structure of source signals but conducted the batch learning and without considering the temporal structure of mixing system. The online VB-ICA (OVB-ICA) [24] performed the online VB learning where the time-varying source signals were characterized but the mixing system was assumed to be fixed. The nonstationary ICA (NS-ICA) [18] performed an online learning of time-varying mixing system. No online tracking of source signals was done. The SMC-ICA [1], [6] conducted the particle filtering and online Bayesian learning via sequential importance sampling. The temporally correlated source signals and mixing matrix were not investigated. However, the proposed NB-ICA and OLGP-ICA methods perform the online learning strategy and deal with the temporally correlated source signals and mixing coefficients. The nonstationary source signals and mixing coefficients are compensated simultaneously.

In implementation of online ICAs, including NS-ICA, SMC-ICA, NB-ICA, and OLGP-ICA, the frame size was fixed to be 0.25 s. The prediction order was set at $p = 6$ when implementing GP-ICA and OLGP-ICA. In the following evaluation, the demixed signals, the ARD parameters, and the mixing matrix were realized from the corresponding variational parameters or the modes of the variational distributions. The computation times of running MATLAB codes of different methods were investigated by a personal computer with Intel Core 2 Duo 2.4-GHz CPU and 4-GB RAM. In our investigation, the computation times for the example set of mixed signals were measured as 2.1, 2.8, 4.7, 4.1, and 5.8 min by using sequential ICAs, including OVB-ICA, NS-ICA, SMC-ICA, NB-ICA, and OLGP-ICA, respectively. SMC-ICA is computationally expensive because of the sampling process and calculation of posterior distributions and importance weights. The additional computation of running OLGP-ICA was caused by the implementation of GP for mixing coefficients and source signals.

B. Effects of the ARD Parameter and Mixing Matrix

First of all, we investigated the effect of compensation of the ARD parameter in NB-ICA method by $\alpha_m^{(l)}\mathbf{I}_N \rightarrow \alpha_m^{(l)}\mathbf{H}_m^{(l)}$ as given in (13) and (14). This compensation attempts to strengthen the effectiveness when generating the mixing matrix $\mathbf{A}^{(l)}$ for nonstationary source separation. Fig. 6 shows the negative-free energy calculated by the VB procedure of NB-ICA algorithm. The same mixed audio signals in Fig. 5(a) are used. The results with and without transformation matrix $\mathbf{H}_m^{(l)}$ were compared. Higher negative-free energy implies better goodness-of-fit between model and demixed signals. The energy was elevated by applying the compensation scheme via a full precision matrix for mixture coefficients. This scheme provided a good prediction of the mixing matrix for the next learning epoch. Fig. 7 displays two diagonal components of $\alpha_m^{(l)}\mathbf{H}_m^{(l)}$ (or ARD parameters) estimated from the same mixed signals. The estimated parameters reflected the activity of latent sources at different time frames. The irrelevant sources were deemphasized. The ARD parameters did effectively indicate the activities of time series of source signals. During the period between 1.5 s and 2.5 s, the first source had a silence segment and was clearly reflected by the ARD parameter of the first demixed signal. During the period between 2.5 s and 3.5 s, the second source had a silence segment and was reflected by the ARD of the second demixed signal. The activity of sources was detected. In addition, Fig. 8 shows the square error between the true mixing matrix $\mathbf{A}_t$ and the estimated mixing matrix $\mathbf{A}_t^{(l)}$ by using NS-ICA, SMC-ICA, NB-ICA, and OLGP-ICA. The square error is accumulated over different mixing coefficients and different samples in a frame. NB-ICA and OLGP-ICA obtained lower square errors than NS-ICA and SMC-ICA at different frames. The OLGP-ICA had more accurate $\mathbf{A}_t^{(l)}$ than NB-ICA at most frames.

C. Evaluation of Signal-to-Interference Ratios

The waveforms of the demixed signals using NB-ICA and OLGP-ICA algorithms are displayed in Fig. 5(b). In this example, the demixed signals using OLGP-ICA are better than those using NB-ICA. The waveforms of source signals, mixed signals, and the demixed signals using

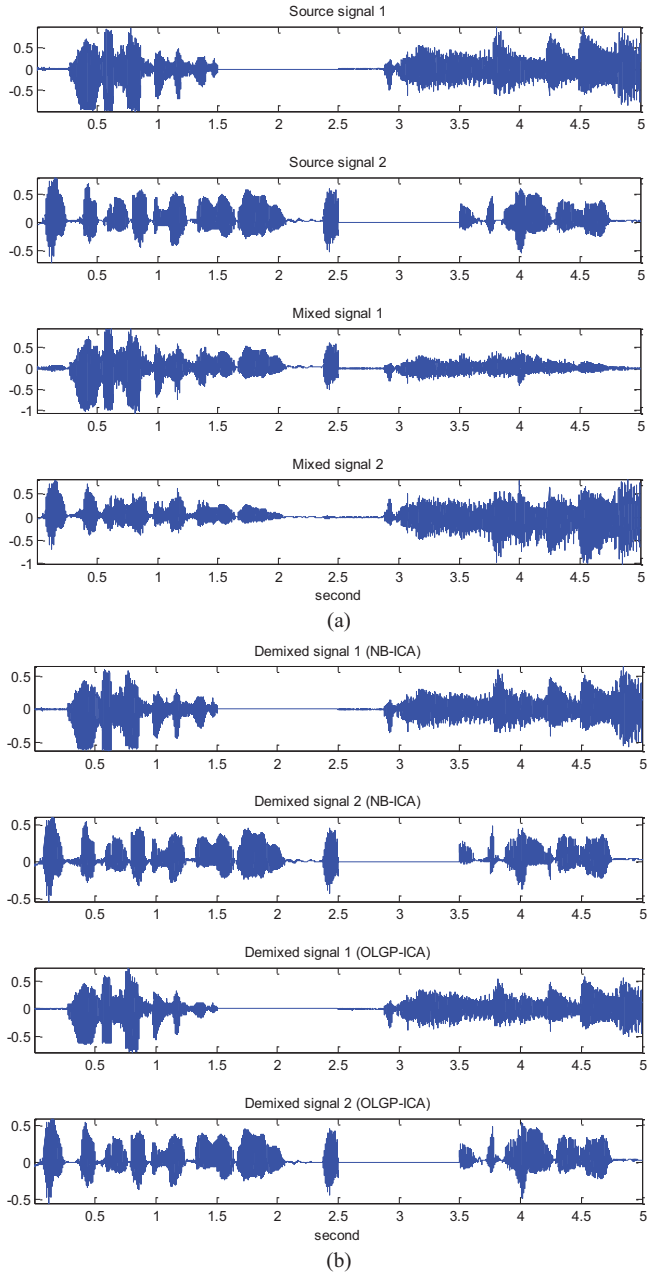

Fig. 5. Waveforms of (a) source signals and mixed signals and (b) demixed signals by using NB-ICA and OLGP-ICA algorithms.

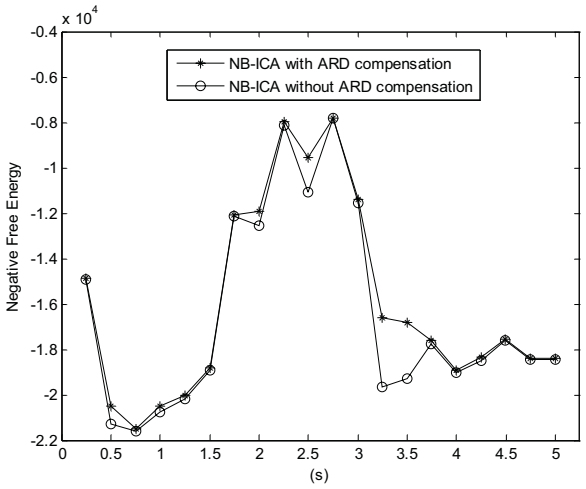

Fig. 6. Comparison of the negative-free energy by using NB-ICA with and without compensation of ARD parameter.

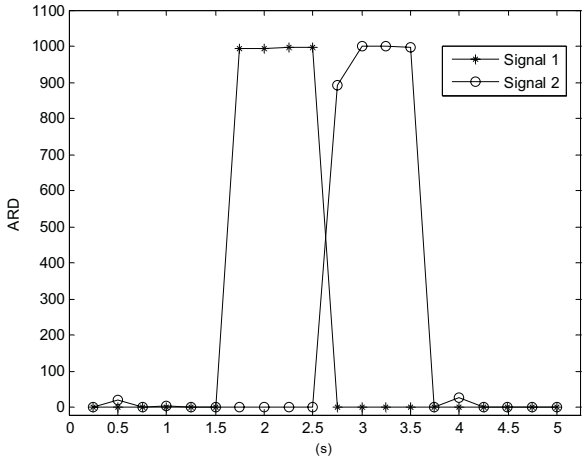

Fig. 7. Comparison of the ARD parameters of the first and the second source signals estimated by using NB-ICA.

VB-ICA, S-ICA, BICA-HMM, GP-ICA, OVB-ICA, NS-ICA, SMC-ICA, NB-ICA, and OLGP-ICA are accessible at: http://chien.cm.nctu.edu.tw/~nb-olgp-ica. To conduct a statistically meaningful evaluation, we further prepared five other sets of mixed signals by applying the same scenarios as mentioned in Section V-A but using different speech signals, music signals, and mixing coefficients $\{a_{nm,t}\}$ with different changing frequencies f_1 and f_2. The length of these test signals was 5 s in average. The investigation over six various sets of mixed signal is performed. A quantitative comparison over different ICA methods is conducted by measuring the signal-to-interference ratios (SIRs) in decibels for all samples in different frames of source signals $\mathbf{s}_m = \{s_{m,t}\}$ and demixed signals $\hat{\mathbf{s}}_m = \{\hat{s}_{m,t}\}$ by using

$$\text{SIR(db)} = 10\log_{10}\frac{\sum_t \|s_{m,t}\|^2}{\sum_t \|\hat{s}_{m,t} - s_{m,t}\|^2} \tag{42}$$

and is reported in Fig. 9. The SIRs are averaged over six test sets. In this comparison, VB-ICA had the worst performance of SIRs because VB-ICA did not deal with nonstationary mixing problems. BICA-HMM and S-ICA presented the solutions to nonstationary source separation. Consistent with the result in [23], S-ICA performed better than BICA-HMM. This is because that S-ICA effectively recovers the source signals that abruptly appear or disappear. GP-ICA obtained higher SIRs than BICA-HMM and S-ICA due to the modeling of temporal correlation, which worked well but with higher computational cost. The performance of using online learning (OVB-ICA, NS-ICA, SMC-ICA, NB-ICA, and OLGP-ICA) was better than that of batch learning (VB-ICA, BICA-HMM, S-ICA, and GP-ICA). The proposed NB-ICA and OLGP-ICA attained higher SIRs than the other ICAs. Among these ICAs, the highest SIRs were achieved by OLGP-ICA.

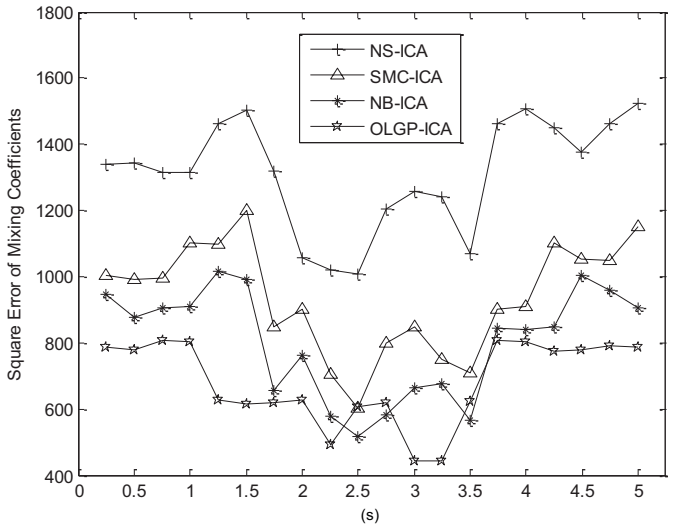

Fig. 8. Comparison of the square errors between the true and the estimated mixing coefficients by using NS-ICA, SMC-ICA, NB-ICA, and OLGP-ICA.

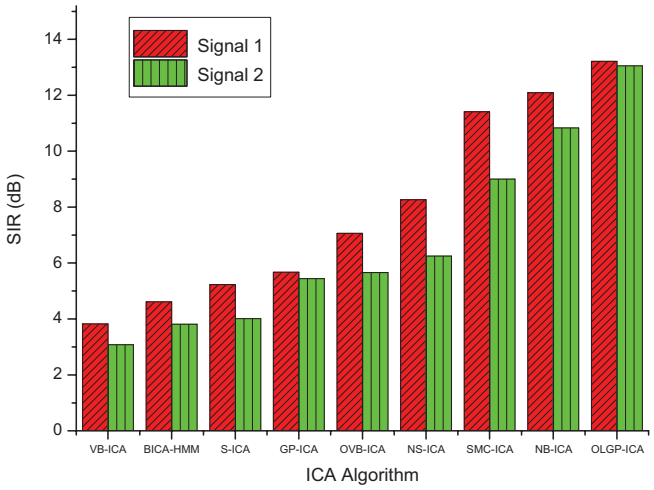

Fig. 9. Comparison of the SIRs of demixed signals by using different ICA methods.

D. Evaluation of Signal Predictability

The audio source signals are temporally correlated. The mixture of these source signals is even more complex than its constituent source signals. The BSS problem was treated by seeking the minimally complex source signals obtained from a set of signal mixtures [46]. It is meaningful to evaluate the performance of demixed signals according to the signal complexity, which is analogous to the contrast function of non-Gaussianity or independence in standard ICA. A simple measure of complexity is formulated in terms of temporal predictability. If a signal value is easy to predict on the basis of previous signal values, that signal has high predictability or equivalently low complexity. The temporal structure can be quantitatively evaluated by the predictability of a signal $\mathbf{s}_m = \{s_{m,t}\}$ measured by [46]

$$F(\mathbf{s}_m) = \log \frac{\sum_t (s_{m,t} - \bar{s}_m)^2}{\sum_t (s_{m,t} - \tilde{s}_{m,t})^2} \tag{43}$$

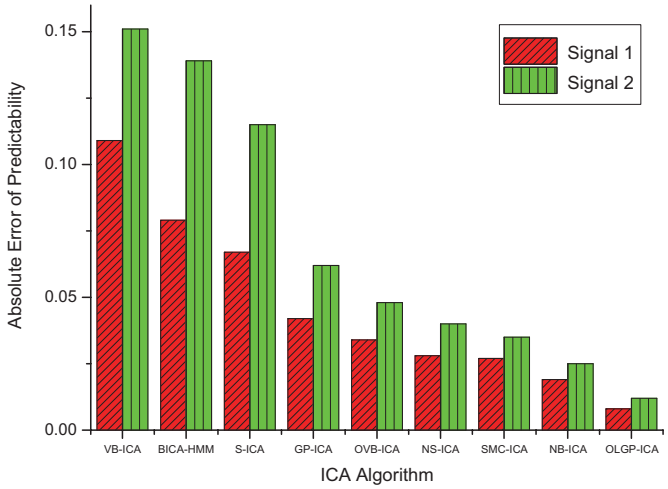

Fig. 10. Comparison of the absolute errors of the predictability between source signals and demixed signals by using different ICA methods.

where $\bar{s}_m = \bar{\eta}\bar{s}_m + (1-\bar{\eta})s_{m,t}$ with $\bar{\eta} = 0.99$ and $\tilde{s}_{m,t} = \tilde{\eta}\tilde{s}_{m,t} + (1-\tilde{\eta})s_{m,t}$ with $\tilde{\eta} = 0.5$. The numerator is a measure of overall variance of extracted signal. The denominator reflects the extent to which $s_{m,t}$ is predicted by a short-term moving average $\tilde{s}_{m,t}$ of previous values in $\mathbf{s}_m$. The demixed signals have high predictability when a large overall signal variance in numerator and a low prediction error in a smooth signal in denominator are obtained. Fig. 10 compares the absolute error of the predictability $|F(\mathbf{s}_m) - F(\hat{\mathbf{s}}_m)|$ between source signals $\mathbf{s}_m$ and the demixed signals $\hat{\mathbf{s}}_m$ estimated by using different ICA methods. The errors of two demixed signals are calculated and averaged over six test signal sets. The lower error implies the better performance that the predictability of demixed signals is closer to that of source signals. Similar to the results of SIRs, the online learning methods using OVB-ICA, NS-ICA, SMC-ICA, NB-ICA, and OLGP-ICA perform better than the batch learning methods using VB-ICA, BICA-HMM, S-ICA, and GP-ICA in terms of signal predictability. Among different ICA methods, the lowest predictability error is achieved by OLGP-ICA. This is because OLGP-ICA does not only deal with the issue of complicated scenarios in nonstationary source separation but also characterizes the temporal correlation existing in source signals and mixing coefficients.

VI. Conclusion

This paper presented two Bayesian learning algorithms NB-ICA and OLGP-ICA for dynamic source separation under nonstationary scenarios and environments. These algorithms were developed by jointly tackling the nonstationary and temporally correlated mixing coefficients and source signals through the online GP. The online learning was built based on a noisy ICA model according to a recursive Bayesian formula based on a reproducible prior/posterior distribution pair. The proposed algorithms adaptively captured the statistics of source signals and the activities of individual sources. The NB-ICA employed an ARD model and efficiently characterized the latent sources by online learning. The sequential

inference procedure was exploited to infer the variational parameters corresponding to different model parameters. The estimated variational parameters served as new hyperparameters of prior density for the next learning epoch when applying a new frame of mixed signals. The demixed signals, mixing matrix, and ARD parameter were realized from the corresponding variational distributions. In addition, the temporal structures of the mixing matrix and source signals were characterized by the GP, and the GP priors were incrementally traced through the reproducible Gaussian prior and posterior distributions. The kernel parameters in GP priors were also estimated in the OLGP-ICA algorithm frame-by-frame. The NB-ICA, OLGP-ICA, and SMC-ICA had comparable model structures and frame-by-frame updating equations. SMC-ICA conducted the frame-by-frame updating of the importance distributions, which were applied to find particles and estimate the SMC-ICA parameters or source signals in sequential importance sampling procedure. Experimental results on source separation of audio signals confirmed the effectiveness of compensating the ARD parameter and the mixing coefficients. The proposed NB-ICA and OLGP-ICA performed better than SMC-ICA and other ICA methods in terms of signal-to-interference ratio and signal predictability. The OLGP-ICA dealt with the most complicated scenario and achieved the best performance among different ICA methods. In the future, the evaluation of different numbers of sources and mixtures shall be investigated. The proposed sequential ICA approaches shall be extended for convolutive and underdetermined BSS.

APPENDIX: SEQUENTIAL VB INFERENCE FOR OLGP-ICA

Continuing from Section IV-B, the hyperparameters $\{\mathbf{M}_s^{(l)}, \mathbf{R}_s^{(l)}\}$ at each frame l are inferred via finding the optimal variational distribution $\hat{q}(\mathbf{s}_m^{(l)})$ of the mth source signal by

$$\begin{aligned}\log \hat{q}(\mathbf{s}_m^{(l)}) \propto\ & E_{q(\Theta\neq S)}[\log p(\mathbf{X}^{(l)}|\mathbf{a}_{nm}^{(l)}, \mathbf{s}_m^{(l)}, \boldsymbol{\varepsilon}_n^{(l)})] \\ & + \log p(\mathbf{s}_m^{(l)}|\boldsymbol{\mu}_{s_m}^{(l-1)}, \mathbf{R}_{s_m}^{(l-1)})\end{aligned} \tag{44}$$

where the expectation is operated over the variational distributions $q(\boldsymbol{\Theta}^{(l)} \neq \mathbf{S}^{(l)})$ by excluding that of $\mathbf{S}^{(l)}$. To combine two terms in RHS of (44), we arrange the first term as a logarithm of Gaussian distribution of a new transformed observation vector $(\boldsymbol{\Psi}_{s_m}^{(l)})^{-1}\tilde{\mathbf{x}}_{s_m}^{(l)}$ with mean $\mathbf{s}_m^{(l)}$ in a form of

$$\begin{aligned}&\exp\{E_{q(\Theta\neq S)}[\log p(\mathbf{X}^{(l)}|\mathbf{a}_{nm}^{(l)}, \mathbf{s}_m^{(l)}, \boldsymbol{\varepsilon}_n^{(l)})]\} \\ &\quad \propto \mathcal{N}((\boldsymbol{\Psi}_{s_m}^{(l)})^{-1}\tilde{\mathbf{x}}_{s_m}^{(l)}|\mathbf{s}_m^{(l)}, (\boldsymbol{\Psi}_{s_m}^{(l)})^{-1})\end{aligned} \tag{45}$$

where $\tilde{\mathbf{x}}_{s_m}^{(l)}$ denotes an $L \times 1$ vector with the tth entry

$$\tilde{x}_{s_m,t}^{(l)} = \sum_{n=1}^{N}\left[\begin{array}{c} E_{q(\Theta\neq S)}[\beta_{n,t}^{(l)}]E_{q(\Theta\neq S)}[a_{nm,t}^{(l)}] \\ \times (x_{n,t}^{(l)} - \sum_{k\neq m}^{M} E_{q(\Theta\neq S)}[a_{nk,t}^{(l)}]E_{q(\Theta\neq S)}[s_{k,t}^{(l)}]) \end{array}\right] \tag{46}$$

and $\boldsymbol{\Psi}_{s_m}^{(l)}$ denotes a diagonal matrix with the tth entry

$$[\boldsymbol{\Psi}_{s_m}^{(l)}]_{tt} = \sum_{n=1}^{N} E_{q(\Theta\neq S)}[\beta_{n,t}^{(l)}]E_{q(\Theta\neq S)}[(a_{nm,t}^{(l)})^2]. \tag{47}$$

By combining two quadratic functions of $\mathbf{s}_m^{(l)}$ in RHS of (44), the variational distribution $\hat{q}(\mathbf{s}_m^{(l)}|\boldsymbol{\mu}_{s_m}^{(l)}, \mathbf{R}_{s_m}^{(l)})$ turns out to be a *new Gaussian posterior distribution* with the updated hyperparameters $\{\boldsymbol{\mu}_{s_m}^{(l)}, \mathbf{R}_{s_m}^{(l)}\}$, which are expressed in the same formulas as (32) and (33) except that the subscripts are all altered by $a_{nm} \to s_m$. Again, the GP prior acts as the conjugate prior so that the reproducible prior/posterior distribution pair is established for sequential VB learning.

In addition, the kernel parameters $\{\boldsymbol{\Lambda}_a^{(l)}, \boldsymbol{\Xi}_a^{(l)}, \boldsymbol{\lambda}_s^{(l)}, \boldsymbol{\xi}_s^{(l)}\}$ in calculation of GP priors are inferred in the OLGP-ICA procedure. The solution to new parameters $\boldsymbol{\Lambda}_a^{(l-1)} = \{\lambda_{a_{nm}}^{(l-1)}\} \to \boldsymbol{\Lambda}_a^{(l)} = \{\lambda_{a_{nm}}^{(l)}\}$ of mixing coefficients $\mathbf{a}_{nm}^{(l)}$ is presented. The individual parameter $\lambda_{a_{nm}}^{(l)}$ is estimated via VB algorithm by maximizing the lower bound of log marginal likelihood

$$\begin{aligned}&\int q(\mathbf{a}_{nm}^{(l)}) \log \left[\begin{array}{c}\exp\{E_{q(\Theta\neq A)}[\log p(\mathbf{X}^{(l)}|\mathbf{a}_{nm}^{(l)}, \mathbf{s}_m^{(l)}, \boldsymbol{\varepsilon}_n^{(l)})]\} \\ \times p(\mathbf{a}_{nm}^{(l)}|\lambda_{a_{nm}}^{(l-1)}, \xi_{a_{nm}}^{(l-1)})(q(\mathbf{a}_{nm}^{(l)}))^{-1}\end{array}\right] d\mathbf{a}_{nm}^{(l)} \\ &= \log \int \exp\{E_{q(\Theta\neq A)}[\log p(\mathbf{X}^{(l)}|\mathbf{a}_{nm}^{(l)}, \mathbf{s}_m^{(l)}, \boldsymbol{\varepsilon}_n^{(l)})]\} \\ &\quad \times p(\mathbf{a}_{nm}^{(l)}|\lambda_{a_{nm}}^{(l-1)}, \xi_{a_{nm}}^{(l-1)}) d\mathbf{a}_{nm}^{(l)} \\ &\propto \log \mathcal{N}((\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1}\tilde{\mathbf{x}}_{a_{nm}}^{(l)}|0, (\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1} + \mathbf{R}_{a_{nm}}^{(l-1)}) \\ &\propto -\frac{1}{2}\log|(\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1} + \mathbf{R}_{a_{nm}}^{(l-1)}| - \frac{1}{2}(\tilde{\mathbf{x}}_{a_{nm}}^{(l)})^T(\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1} \\ &\quad \times((\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1} + \mathbf{R}_{a_{nm}}^{(l-1)})^{-1}(\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1}\tilde{\mathbf{x}}_{a_{nm}}^{(l)}.\end{aligned} \tag{48}$$

In (48), the RHS is obtained by substituting into the normalized variational distribution $q(\mathbf{a}_{nm}^{(l)})$ as determined in (28). The kernel parameter $\lambda_{a_{nm}}^{(l-1)}$ exists in the covariance matrix $\mathbf{R}_{a_{nm}}^{(l-1)} = \mathbf{K}_{a_{nm}}^{(l-1)}$ of GP prior $p(\mathbf{a}_{nm}^{(l)}|\boldsymbol{\mu}_{a_{nm}}^{(l-1)}, \mathbf{R}_{a_{nm}}^{(l-1)})$. The integral is operated over a Gaussian distribution of $\mathbf{a}_{nm}^{(l)}$ and comes up with a new Gaussian distribution. However, there is no closed-form solution to this optimal hyperparameter. The steepest descent algorithm is applied to find the solution by using a learning rate and the differentiation of (48) with respect to $\lambda_{a_{nm}}^{(l-1)}$, which is written by [32]

$$\begin{aligned}&-\frac{1}{2}\mathrm{Tr}\left[((\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1} + \mathbf{R}_{a_{nm}}^{(l-1)})^{-1}\frac{\partial \mathbf{R}_{a_{nm}}^{(l-1)}}{\partial \lambda_{a_{nm}}^{(l-1)}}\right] \\ &\quad + \frac{1}{2}(\tilde{\mathbf{x}}_{a_{nm}}^{(l)})^T(\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1}((\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1} + \mathbf{R}_{a_{nm}}^{(l-1)})^{-1} \\ &\quad \times \frac{\partial \mathbf{R}_{a_{nm}}^{(l-1)}}{\partial \lambda_{a_{nm}}^{(l-1)}}((\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1} + \mathbf{R}_{a_{nm}}^{(l-1)})^{-1}(\boldsymbol{\Psi}_{a_{nm}}^{(l)})^{-1}\tilde{\mathbf{x}}_{a_{nm}}^{(l)}.\end{aligned} \tag{49}$$

The solutions to the other hyperparameters $\{\xi_{a_{nm}}^{(l-1)}, \lambda_{s_m}^{(l-1)}, \xi_{s_m}^{(l-1)}\} \to \{\xi_{a_{nm}}^{(l)}, \lambda_{s_m}^{(l)}, \xi_{s_m}^{(l)}\}$ are similarly derived by applying the steepest descent algorithms and considering the objective function in (48).

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive feedback and helpful suggestions.

REFERENCES

[1] A. Ahmed, C. Andrieu, A. Doucet, and P. J. W. Rayner, "On-line non-stationary ICA using mixture models," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, vol. 5. Jun. 2000, pp. 3148–3151.

[2] M. Anderson, T. Adali, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1672–1683, Apr. 2012.

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

[4] W. Bian and X. Chen, "Smoothing neural network for constrained non-Lipschitz optimization with applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 399–411, Mar. 2012.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, USA: Springer-Verlag, 2006.

[6] M. Costagli and E. E. Kuruoglu, "Image separation using particle filters," *Digit. Signal Process.*, vol. 17, no. 5, pp. 935–946, 2007.

[7] A. T. Cemgil, C. Fevotte, and S. J. Godsill, "Variational and stochastic inference for Bayesian source separation," *Digit. Signal Process.*, vol. 17, no. 5, pp. 891-913, 2007.

[8] K. Chan, T. W. Lee, and T. J. Sejnowski, "Variational learning of clusters of undercomplete nonsymmetric independent components," *J. Mach. Learn. Res.*, vol. 3, pp. 99–114, Aug. 2002.

[9] J.-T. Chien, "Online hierarchical transformation of hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 656–667, Nov. 1999.

[10] J.-T. Chien and B.-C. Chen, "A new independent component analysis for speech recognition and separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1245–1254, Jul. 2006.

[11] J.-T. Chien and M.-S. Wu, "Adaptive Bayesian latent semantic analysis," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 1, pp. 198–207, Jan. 2008.

[12] J.-T. Chien and J.-C. Chen, "Recursive Bayesian linear regression for adaptive classification," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 565–575, Feb. 2009.

[13] J.-T. Chien and H.-L. Hsieh, "Convex divergence ICA for blind source separation," *IEEE Trans. Audio, Speech Language Process.*, vol. 20, no. 1, pp. 302–313, Jan. 2012.

[14] J.-T. Chien and H.-L. Hsieh, "Bayesian group sparse learning for nonnegative matrix factorization," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2012.

[15] R. A. Choudrey and S. J. Roberts, "Bayesian ICA with hidden Markov sources," in *Proc. Int. Workshop Independ. Compon. Anal. Blind Signal Separat.*, 2003, pp. 809–814.

[16] R. Choudrey, W. D. Penny, and S. J. Roberts, "An ensemble learning approach to independent component analysis," in *Proc. IEEE Workshop Neural Netw. Signal Process.*, 2000, pp. 435–444.

[17] A. Doucet, N. J. Gordon, and V. Krishnamurthy, "Particle filters for state estimation of jump Markov linear systems," *IEEE Trans. Signal Process.*, vol. 49, no. 3, pp. 613–624, Mar. 2001.

[18] R. Everson and S. Roberts, "Blind source separation for non-stationary mixing," *J. VLSI Signal Process.*, vol. 26, nos. 1–2, pp. 15–23, 2000.

[19] B. Gao, W. L. Woo, and S. S. Dlay, "Adaptive sparsity non-negative matrix factorization for single-channel source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 989–1001, Sep. 2011.

[20] B. Gao, W. L. Woo, and S. S. Dlay, "Single-channel source separation using EMD-subband variable regularized sparse features," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 961–976, May 2011.

[21] B. Gao, W. L. Woo, and S. S. Dlay, "Variational regularized 2-D nonnegative matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 703–716, May 2012.

[22] D. Gencaga, E. E. Kuruoglu, and A. Ertuzun, "Modeling non-Gaussian time-varying vector autoregressive processes by particle filtering," *Multidimens. Syst. Signal Process.*, vol. 21, no. 1, pp. 73–85, 2010.

[23] J. Hirayama, S. Maeda, and S. Ishii, "Markov and semi-Markov switching of source appearances for nonstationary independent component analysis," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1326–1342, Sep. 2007.

[24] A. Honkela and H. Valpola, "On-line variational Bayesian learning," in *Proc. Int. Workshop Independ. Compon. Anal. Blind Signal Separat.*, 2003, pp. 803–808.

[25] H.-L. Hsieh and J.-T. Chien, "Online Bayesian learning for dynamic source separation," in *Proc. Int. Conf. Acous., Speech Signal Process.*, 2010, pp. 1950–1953.

[26] H.-L. Hsieh and J.-T. Chien, "Online Gaussian process for nonstationary speech separation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 394–397.

[27] Q. Huang, J. Yang, and S. Wei, "Temporally correlated source separation using variational Bayesian learning approach," *Digit. Signal Process.*, vol. 17, no. 5, pp. 873–890, 2007.

[28] Q. Huang, J. Yang, and Y. Zhou, "Bayesian nonstationary source separation," *Neurocomputing*, vol. 71, nos. 7–9, pp. 1714–1729, 2008.

[29] M. M. Ichir and A. Mohammad-Djafari, "Hidden Markov models for wavelet-based blind source separation," *IEEE Trans. Image Process.*, vol. 15, no. 7, pp. 1887–1899, Jul. 2006.

[30] Z. Koldovsky, J. Malek, P. Tichavsky, Y. Deville, and S. Hosseini, "Blind separation of piecewise stationary non-Gaussian sources," *Signal Process.*, vol. 89, no. 12, pp. 2570–2584, 2009.

[31] N. D. Lawrence and C. M. Bishop, "Variational Bayesian independent component analysis," Univ. Cambridge, Cambridge, U.K., Tech. Rep., 2000.

[32] J. Luttinen and A. Ilin, "Variational Gaussian-process factor analysis for modeling spatiotemporal data," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2009, pp. 1177–1185.

[33] D. J. C. MacKay, "Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks," *Netw., Comput. Neural Syst.*, vol. 6, no. 3, pp. 469-505, 1995.

[34] J. W. Miskin, "Ensemble learning for independent component analysis," Ph.D. dissertation, Dept. Archit., Univ. Cambridge, Cambridge, U.K. 2000.

[35] A. Mohammad-Djafari and K. H. Knuth, "Bayesian approaches," in *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, P. Comon and C. Jutten, Eds. New York, USA: Elsevier, 2010, pp. 467–513.

[36] M. K. I. Molla and K. Hirose, "Single-mixture audio source separation by subspace decomposition of Hilbert spectrum," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 3, pp. 893–900, Mar. 2007.

[37] 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 Process.*, vol. 54, no. 11, pp. 4133–4145, Nov. 2006.

[38] S. M. Naqvi, Y. Zhang, and J. A. Chambers, "Multimodal blind source separation for moving sources," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 125–128.

[39] S. Park and S. Choi, "Gaussian processes for source separation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 1909–1912.

[40] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Cambridge, MA, USA: MIT Press, 2006.

[41] D. B. Rowe, *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*, Boca Raton, FL, USA: CRC Press, 2003.

[42] M. N. Schmidt and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proc. Int. Workshop Independ. Compon. Anal. Blind Signal Separat.*, 2006, pp. 700–707.

[43] H. Snoussi and A. Mohammad-Djafari, "Bayesian unsupervised learning for source separation with mixture of Gaussian prior," *J. VLSI Signal Process.*, vol. 37, nos. 2–3, pp. 263–279, 2004.

[44] H. Snoussi and A. Mohammad-Djafari, "Fast joint separation and segmentation of mixed images," *J. Electron. Imag.*, vol. 13, no. 2, pp. 349–361, 2004.

[45] J. Spragins, "A note on the iterative application of Bayes' rule," *IEEE Trans. Inf. Theory*, vol. 11, no. 4, pp. 544–549, Oct. 1965.

[46] J. V. Stone, "Blind source separation using temporal predictability," *Neurocomputation*, vol. 13, no. 7, pp. 1559–1574, 2001.

[47] A. Stuhlsatz, J. Lippel, and T. Zielke, "Feature extraction with deep neural networks by a generalized discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 4, pp. 596–608, Apr. 2012.

[48] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jan. 2001.

[49] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[50] S. Xie, L. Yang, J.-M. Yang, G. Zhou, and Y. Xiang, "Time-frequency approach to underdetermined blind source separation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 2, pp. 306–316, Feb. 2012.

[51] Z. Yang, Y. Xiang, S. Xie, S. Ding, and Y. Rong, "Nonnegative blind source separation by sparse component analysis based on determinant measure," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 10, pp. 1601–1610, Oct. 2012.

[52] Y. Zhang, P. Liu, J.-T. Chien, and F. Soong, "An evidence framework for Bayesian learning of continuous-density hidden Markov models," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 3857–3860.

Jen-Tzung Chien (S'97–A'98–M'99–SM'04) received the Ph.D. degree from National Tsing Hua University, Hsinchu, Taiwan, in 1997.

He was with National Cheng Kung University, Tainan, Taiwan, from 1997 to 2012. Since 2012, he has been with the Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, where he is currently a Distinguished Professor. He was a Visiting Researcher with Panasonic Technologies Inc., Santa Barbara, CA, USA, the Tokyo Institute of Technology, Tokyo, Japan, the Georgia Institute of Technology, Atlanta, GA, USA, Microsoft Research Asia, Beijing, China, and the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. His current research interests include machine learning, blind source separation, speech recognition, face recognition, and information retrieval.

Dr. Chien was a recipient of the Ta-You Wu Memorial Award from the National Science Council (NSC) of Taiwan in 2003, the Research Award for Junior Research Investigators from Academia Sinica of Taiwan in 2004, the NSC Distinguished Research Award in 2006 and 2010, and the Best Paper Award at the IEEE Automatic Speech Recognition and Understanding Workshop in 2011. He was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS from 2008 to 2011, and a Tutorial Speaker of the ICASSP in 2012, and the APSIPA Distinguished Lecturer from 2012 to 2013.

Hsin-Lung Hsieh received the B.S. and M.S. degrees from I-Shou University, Kaohsiung, Taiwan, in 2001 and 2003, respectively, and the Ph.D. degree from National Cheng Kung University, Tainan, Taiwan, in 2012, all in computer science and information engineering.

His current research interests include machine learning, blind source separation, independent component analysis, and nonnegative matrix factorization.