Bayesian Source Separation

Jen-Tzung Chien

Department of Electrical and Computer Engineering

National Chiao Tung University, Taiwan

July 15, 2015

Outline

- Introduction
- Model-Based Source Separation
- Adaptive Learning Machine
- Case Study: Independent Component Analysis
- Case Study: Nonnegative Matrix Factorization
- Summarization and Future Trend

Outline

- **Introduction**
- Model-Based Source Separation
- Adaptive Learning Machine
- Case Study: Independent Component Analysis
- Case Study: Nonnegative Matrix Factorization
- Summarization and Future Trend

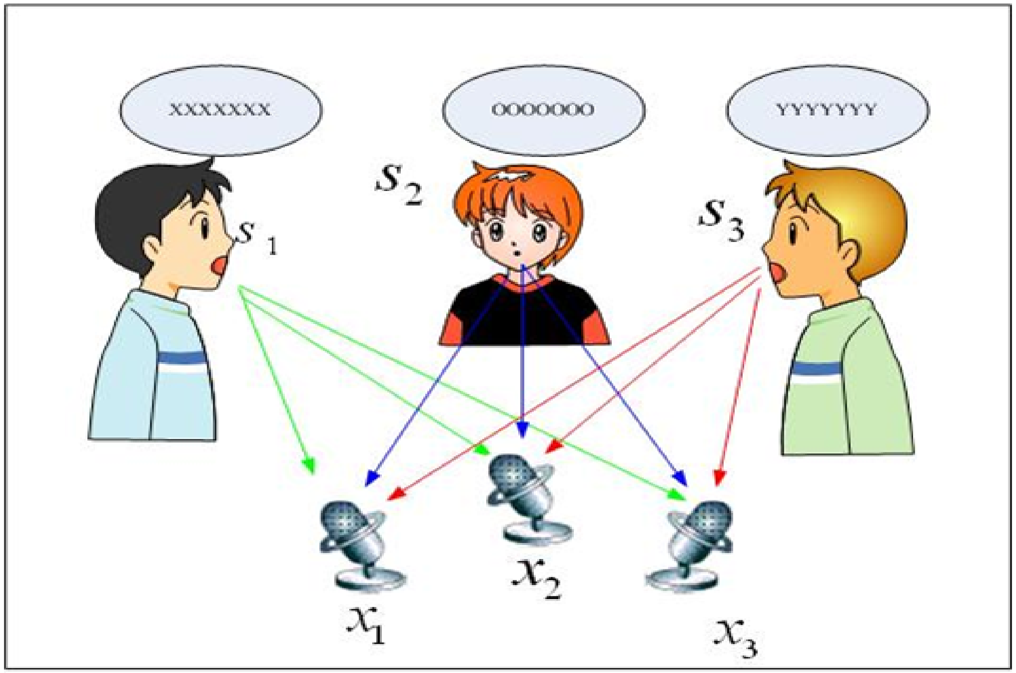
Introduction

- Blind source separation
- Application and challenge
- Overview of this talk

What is BSS?

- Blind source separation (BSS) is to separate a set of source signals from a set of mixed signals, without the aid of information (or with very little information) about the source signals or the mixing process

Cocktail party problem



Linear mixing system

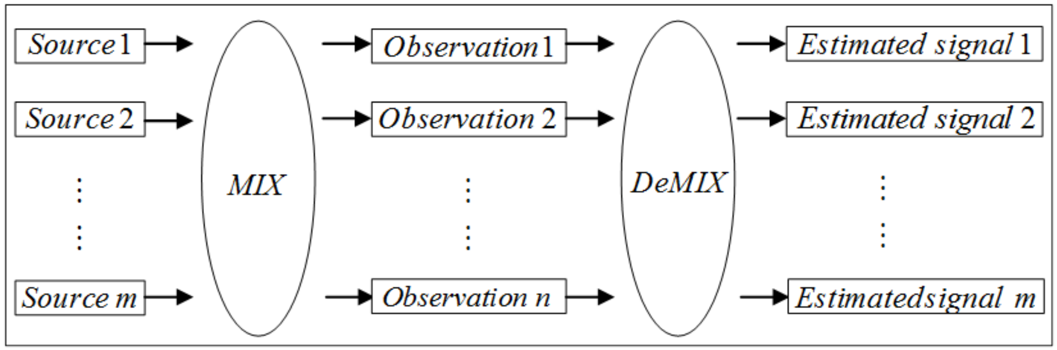
Instantaneous and noiseless mixing system

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t)$$
$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t)$$
$$x_3(t) = a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t)$$

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad \mathbf{W} = \mathbf{A}^{-1}, \quad \mathbf{y} = \mathbf{W}\mathbf{x}$$

- Goal
  - Unknown: $\mathbf{A}$ and $\mathbf{s}$
  - Reconstruct the source signal via demixing matrix $\mathbf{W}$

Linear mixing in general



$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + \cdots + a_{1m}s_m(t)$$

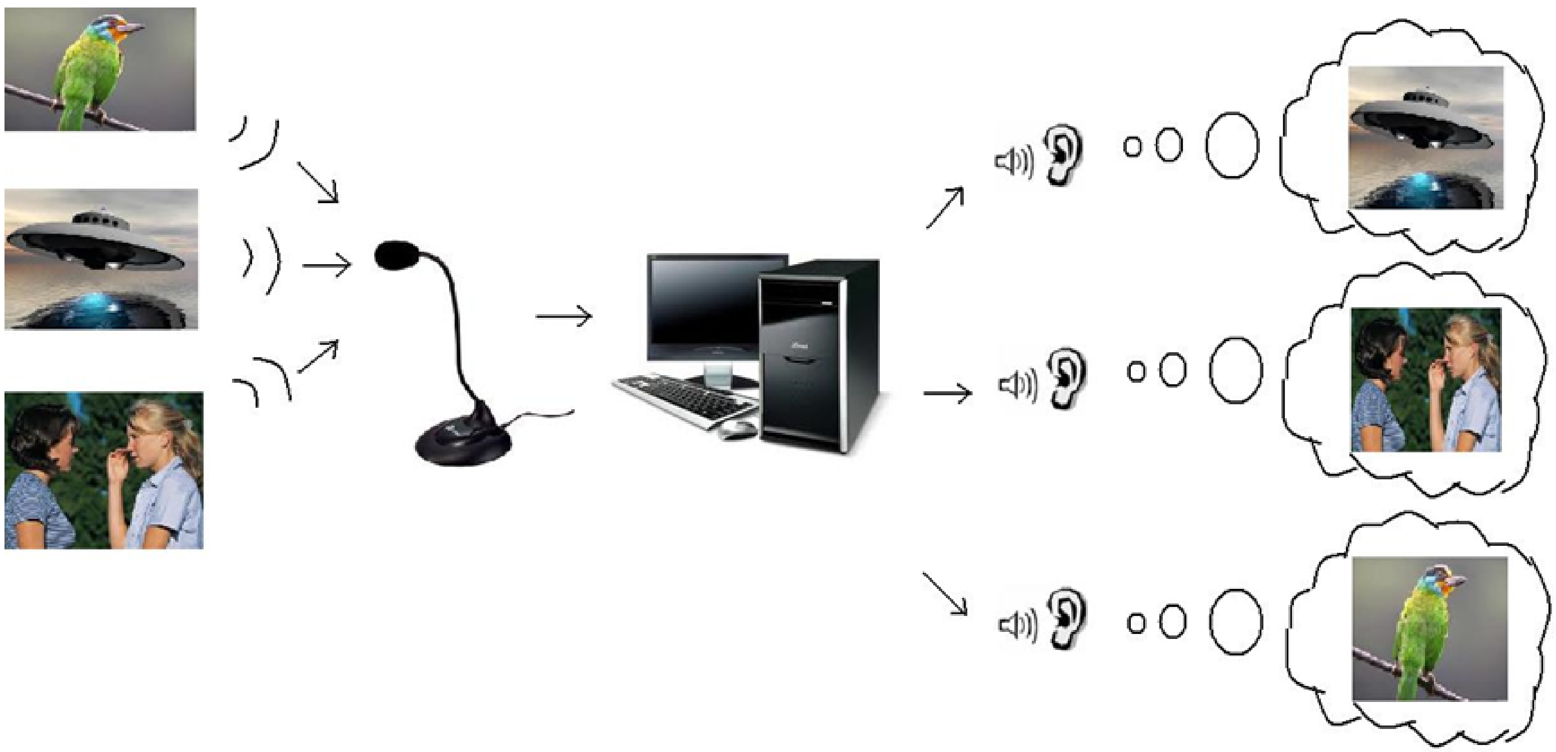
$$\vdots$$

$$x_n(t) = a_{n1}s_1(t) + a_{n2}s_2(t) + \cdots + a_{nm}s_m(t)$$

- Three conditions in multi-channel source separation
  - determined system: $n = m$
  - overdetermined system: $n > m$
  - underdetermined system: $n < m$

Single-channel source separation

- BSS is in general highly underdetermined
- Many applications involve single-channel source separation problem ($n = 1$)



Introduction

- Blind source separation
- Application and challenge
- Overview of this talk

Applications

- Unsupervised learning in general
  - latent component analysis
  - data clustering and mining
- Speech separation
  - speech enhancement, noise reduction
  - teleconferencing, dialogue system
  - hands-free human-machine communication
- Music separation
  - singing-voice separation
  - instrument separation and classification
  - sound classification
  - auditory scene classification
  - music information retrieval

Challenges in audio source separation

- Microphone array signal processing (Benesty et al., 2008)
  - delay-and-sum beamforming
  - denoising, dereverberation, localization
- Convolutive mixtures
  - frequency-domain BSS (Sawada et al., 2007)
- Room reverberation (Yoshioka et al., 2012)
  - teleconferencing, interactive TV, hands-free interface
  - distant-talking speech recognition
- Unknown number of sources (Araki et al., 2009)
  - sparse source separation
  - modeling for direction of arrival

- Unknown model complexity
  - model selection (Fevotte, 2007)
  - model uncertainty
  - unknown number of bases
  - unknown model structure
  - improper model assumption
  - complicated mixing system

- Heterogeneous environments
  - noise contamination
  - adverse condition
  - nonstationary mixing system (Chien and Hsieh, 2013)
  - source is moving
  - source replacement
  - number of sources is changed

Two categories

- Front-end processing
  - adaptive signal processing
  - analysis of information on each source
  - time-frequency modeling and masking
  - identification of mixing system

- Back-end learning
  - adaptive machine learning
  - only using the information about mixture signals
  - model-based approaches
  - statistical model for the whole system
  - inference and learning from a set of samples
  - joint speech separation and recognition (Rennie et al., 2010)

Model-based approach

- Model-based approach aims to incorporate the physical phenomena, measurements, uncertainties and noises in the form of mathematical models
- This approach is developed in a unified manner through different algorithms, examples, applications, and case studies
- Main-stream methods are based on the statistical models
- Machine learning provides a wide range of model-based approaches for blind source separation

Introduction

- Blind source separation
- Application and challenge
- Overview of this talk

Overview of this talk

- Applications
  - speech and music separation
  - instrument separation, singing-voice separation
- Separation models
  - independent component analysis
  - nonstationary Bayesian ICA, online Gaussian process ICA
  - nonnegative matrix factorization - Bayesian NMF, group sparse NMF
- Learning algorithms
  - Bayesian learning, model regularization, structural learning
  - online learning, sparse learning

Outline

- Introduction
- **Model-Based Source Separation**
- Adaptive Learning Machine
- Case Study: Independent Component Analysis
- Case Study: Nonnegative Matrix Factorization
- Summarization and Future Trend

Independent component analysis

- ICA (Comon, 1994) is essential for blind source separation
- ICA is applied to separate the mixed signals and find the independent components
- The demixed components can be grouped into clusters where the intra-cluster elements are dependent and inter-cluster elements are independent
- ICA provides unsupervised learning approach to acoustic modeling, signal separation and many others

Assumptions in ICA

- Three assumptions
  - sources are statistically independent
  - independent component has non-gaussian distribution
  - mixing system is determined, i.e. $n = m \Rightarrow$ square mixing matrix

Linear noiseless ICA: $\mathbf{X} = \mathbf{AS}$

$$\begin{bmatrix} x_{11} & \cdots & x_{1t} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nt} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} s_{11} & \cdots & s_{1t} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nt} \end{bmatrix}$$

ICA learning rule

- ICA demixing matrix can be estimated by optimizing an objective or a contrast function $D(\mathbf{X}, \mathbf{W})$ using a set of samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ via
  - gradient descent algorithm

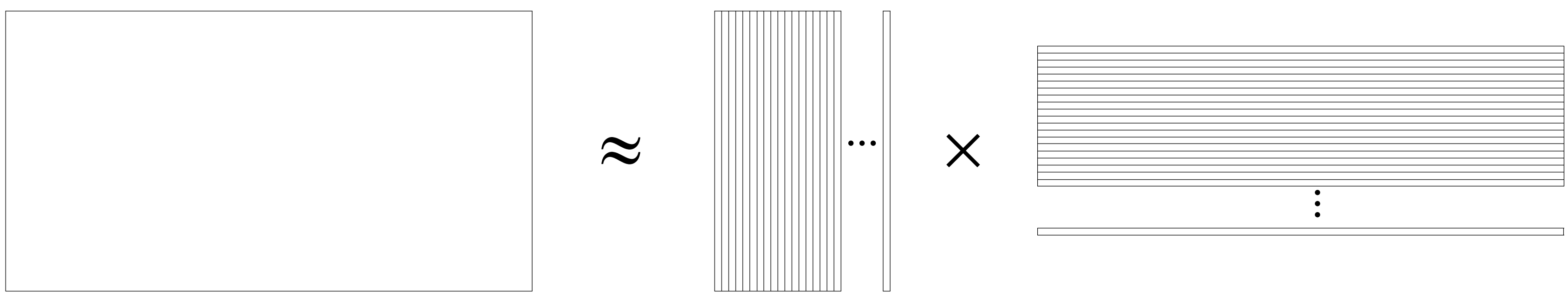
$$\mathbf{W}^{(\tau+1)} = \mathbf{W}^{(\tau)} - \eta \frac{\partial D(\mathbf{X}, \mathbf{W}^{(\tau)})}{\partial \mathbf{W}^{(\tau)}}$$

  - natural gradient algorithm (Amari, 1998)

$$\mathbf{W}^{(\tau+1)} = \mathbf{W}^{(\tau)} - \eta \frac{\partial D(\mathbf{X}, \mathbf{W}^{(\tau)})}{\partial \mathbf{W}^{(\tau)}} (\mathbf{W}^{(\tau)})^T \mathbf{W}^{(\tau)}$$

Nonnegative matrix factorization

$$\mathbf{X} \in \mathcal{R}_+^{M \times N} \qquad \mathbf{B} \in \mathcal{R}_+^{M \times K} \qquad \mathbf{W} \in \mathcal{R}_+^{K \times N}$$



Some properties

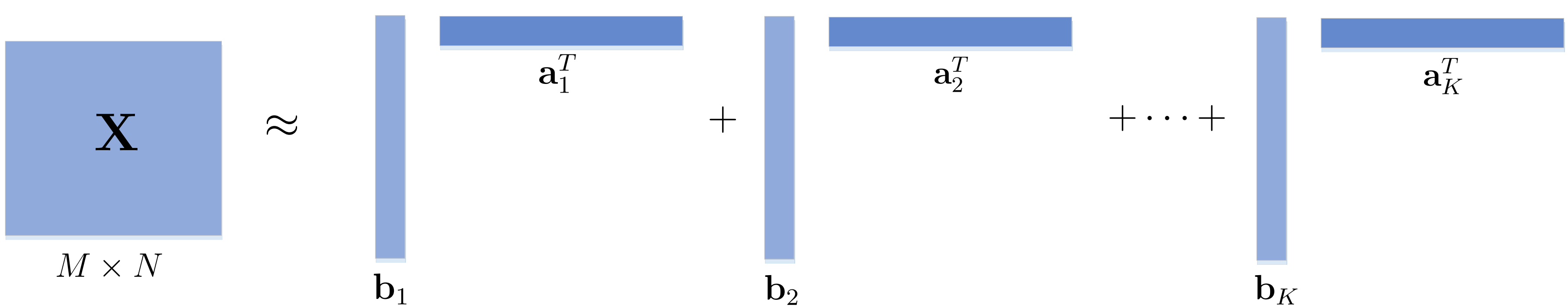
- NMF (Lee and Seung, 1999) conducts the parts-based representation
  - only additive combinations are allowed
  - only a few components are active to encode input data
  - sparsity constraint is imposed
- Nonnegative constraint is imposed to reflect a wide range of nature signals
  - pixel intensities, amplitude spectra, occurrence counts and many others
- NMF does not assume independent sources
- NMF has been popular for single-channel source separation

NMF representation

- NMF aims to decompose the nonnegative data matrix $\mathbf{X} \in \mathcal{R}_+^{M \times N}$ into a product of a nonnegative basis matrix $\mathbf{B} \in \mathcal{R}_+^{M \times K}$ and a nonnegative weight matrix $\mathbf{W} = \mathbf{A}^T = [\mathbf{a}_1, \ldots, \mathbf{a}_K]^T \in \mathcal{R}_+^{K \times N}$

$$\mathbf{X} \approx \mathbf{BW} = \mathbf{BA}^T = \sum_k \mathbf{b}_k \circ \mathbf{a}_k \Rightarrow X_{mn} \approx [\mathbf{BW}]_{mn} = \sum_k B_{mk} W_{kn}$$

- Bilinear NMF: sum of linear combination of rank-one nonnegative matrices



NMF objective function

- Squared Euclidean distance $\Rightarrow$ EU-NMF

$$D_{\mathsf{EU}}(\mathbf{X} \,\|\, \mathbf{BW}) = \sum_{m,n} (X_{mn} - [\mathbf{BW}]_{mn})^2$$

- Kullback-Leibler divergence $\Rightarrow$ KL-NMF

$$D_{\mathsf{KL}}(\mathbf{X} \,\|\, \mathbf{BW}) = \sum_{m,n} \left(X_{mn} \log \frac{X_{mn}}{[\mathbf{BW}]_{mn}} + [\mathbf{BW}]_{mn} - X_{mn} \right)$$

- Itakura-Saito distance $\Rightarrow$ IS-NMF

$$D_{\mathsf{IS}}(\mathbf{X} \,\|\, \mathbf{BW}) = \sum_{m,n} \left(\frac{X_{mn}}{[\mathbf{BW}]_{mn}} - \log \frac{X_{mn}}{[\mathbf{BW}]_{mn}} - 1 \right)$$

Sparsity constraint

- Only a few components are active to handle overcomplete problem
- Objective function with sparsity constraint (Hoyer, 2004)

$$\min_{\mathbf{B},\mathbf{W}\geq 0} D(\mathbf{X} \,\|\, \mathbf{BW}) + \lambda g(\mathbf{W})$$

where $g(\cdot)$ is a penalty function for sparsity control and λ is a regularization parameter

Why nonnegativity and sparsity constraints?

- Many real-word data are nonnegative and the corresponding hidden components have physical meaning only with nonnegativity
- Sparseness is closely related to feature selection
- Nonnegativity relates to probability distribution
- It is important to seek the trade-off between interpretability and statistical fidelity

Multiplicative updating rule

	NMF	Sparse NMF
squared Euclidean distance	$\mathbf{B} \leftarrow \mathbf{B} \odot \frac{\mathbf{X}\mathbf{W}^T}{\mathbf{B}\mathbf{W}\mathbf{W}^T}$ $\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\mathbf{B}^T\mathbf{X}}{\mathbf{B}^T\mathbf{B}\mathbf{W}}$	$\mathbf{B} \leftarrow \mathbf{B} \odot \frac{\mathbf{X}\mathbf{W}^T + \mathbf{B}\odot(\mathbf{1}(\mathbf{B}\mathbf{W}\mathbf{W}^T \odot \mathbf{B}))}{\mathbf{B}\mathbf{W}\mathbf{W}^T + \mathbf{B}\odot(\mathbf{1}(\mathbf{X}\mathbf{W}^T \odot \mathbf{B}))}$ $\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\mathbf{B}^T\mathbf{X}}{\mathbf{B}^T\mathbf{B}\mathbf{W} + \lambda}$
Kullback-Leibler divergence	$\mathbf{B} \leftarrow \mathbf{B} \odot \frac{\frac{\mathbf{X}}{\mathbf{B}\mathbf{W}}\mathbf{W}^T}{\mathbf{1}\mathbf{W}^T}$ $\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\mathbf{B}^T\frac{\mathbf{X}}{\mathbf{B}\mathbf{W}}}{\mathbf{B}^T\mathbf{1}}$	$\mathbf{B} \leftarrow \mathbf{B} \odot \frac{\frac{\mathbf{X}}{\mathbf{B}\mathbf{W}}\mathbf{W}^T + \mathbf{B}\odot(\mathbf{1}(\mathbf{1}\mathbf{W}^T \odot \mathbf{B}))}{\mathbf{1}\mathbf{W}^T + \mathbf{B}\odot(\mathbf{1}(\frac{\mathbf{X}}{\mathbf{B}\mathbf{W}}\mathbf{W}^T \odot \mathbf{B}))}$ $\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\mathbf{B}^T\frac{\mathbf{X}}{\mathbf{B}\mathbf{W}}}{\mathbf{B}^T\mathbf{1} + \lambda}$

Outline

- Introduction
- Model-Based Source Separation
- **Adaptive Learning Machine**
- Case Study: Independent Component Analysis
- Case Study: Nonnegative Matrix Factorization
- Summarization and Future Trend

Adaptive Learning Machine

- Bayesian learning
- Sparse learning
- Online learning

Challenges in model-based approach

- We are facing the challenges of big data
- We need tools for modeling, analyzing, searching, recognizing and understanding real-world data
- Our modeling tools should
  - faithfully represent uncertainty in model structure and its parameters
  - reflect noise condition in observed data
  - be automated and adaptive
  - assure robustness to ill-posed or mismatch condition
  - scalable for large data set
  - deal with over-estimation or under-estimation
- Uncertainty can be properly expressed by prior distribution or process

Bayesian source separation

- Real-world blind source separation
  - unsupervised learning of source signals and mixing process
  - number of sources is unknown
  - underdetermined and sparse sources
  - dynamic time-varying mixing system
  - mixing process is nonstationary

- Why Bayesian? (Fevotte, 2007)
  - automatic relevance determination is used to determine the number of sources
  - recursive Bayesian for online tracking of nonstationary conditions
  - Gaussian process explore the temporal structure of time-varying sources
  - approximate Bayesian inference

Adaptive Learning Machine

- Bayesian learning
- Sparse learning
- Online learning

Sparse coding

- Sparse representation is crucial for blind source separation (Li et al., 2014)
- Sparse coding aims to find a sparse measurement based on a set of over-determined basis vectors
- Basis representation of data $\mathbf{x} \in \mathcal{R}^D$

$$\mathbf{x} = \mathbf{B}\mathbf{w}$$

  - basis vectors or dictionary $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_N]$
  - sensing weights $\mathbf{w} \in \mathcal{R}^N$
  - reconstruction errors $\|\mathbf{x} - \mathbf{B}\mathbf{w}\|_2^2$
- Sensing weights are prone to be sparse in ill-posed conditions

ℓ_1-regularized objective function

- Lasso regularization (Tibshirani, 1996) is imposed to fulfill sparse coding via

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \ \frac{1}{2}\|\mathbf{x} - \mathbf{B}\mathbf{w}\|^2 + \eta\|\mathbf{w}\|_1$$

- A relatively small set of relevant bases is selected to represent target data

- Maximum *a posteriori* (MAP) estimation does the same thing

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \ \{-\log p(\mathbf{x}|\mathbf{w}) - \log p(\mathbf{w})\}$$

  - Gaussian likelihood $p(\mathbf{x}|\mathbf{w}) = \mathcal{N}(\mathbf{x}|\mathbf{B}\mathbf{w}, \mathbf{I})$
  - Laplace prior $p(\mathbf{w}|\eta) = \frac{\eta}{2}\exp(-\eta\|\mathbf{w}\|_1)$

Sparse Bayesian learning

- Bayesian sensing aims to yield the error bars or distribution estimates of the true signals

- Prior density of sensing weights is incorporated

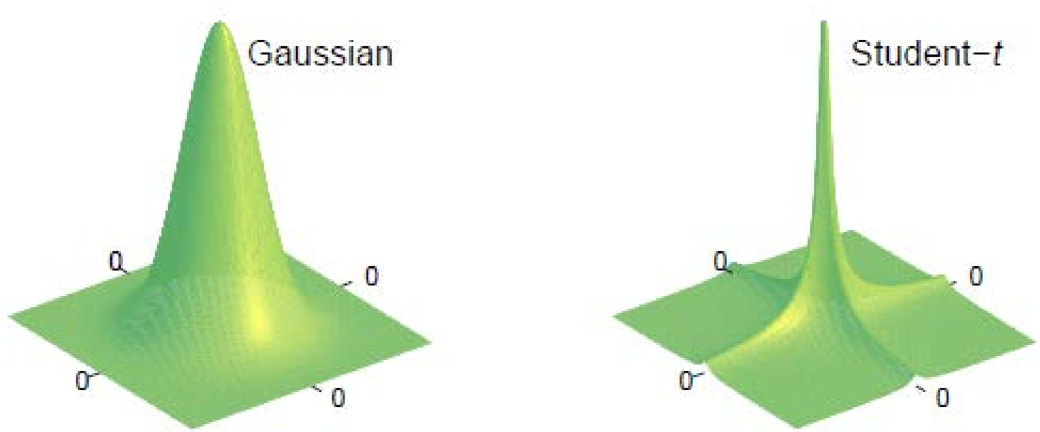
$$p(\mathbf{w}|\mathcal{A}) = \mathcal{N}(\mathbf{w}|0, \mathsf{diag}\{\alpha_n^{-1}\}) = \prod_{n=1}^{N} \mathcal{N}(w_n|0, \alpha_n^{-1})$$

- Automatic relevance determination (ARD) parameter α_n reflects how an observation is relevant to a basis vector (Tipping, 2001)

- If ARD is modeled by a gamma density, the marginal distribution of weights turns out to be an Student's t distribution which is a sparse prior

$$p(\mathbf{w}|a,b) = \prod_{n=1}^{N} \int_0^\infty \mathcal{N}(w_n|0,\alpha_n^{-1})\mathcal{G}(\alpha_n|a,b)d\alpha_n$$

$$\propto \prod_{n=1}^{N} (b + w_n^2/2)^{-(a+1/2)}$$



- Sparse Bayesian learning has been popular for model-based BSS

Adaptive Learning Machine

- Bayesian learning
- Sparse learning
- Online learning

Online learning

- Online learning is preferred when data becomes available in a sequential mode
- Model is updated in a scalable fashion
- Instead of updating model in batch mode using cost function $E = \sum_t E_t$ from all samples $\{\mathbf{x}_t\}$, the online or stochastic learning using gradient descent algorithm is performed according to the cost function from a minibatch or an individual sample E_t

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla E_t$$

- Bayesian theory provides a meaningful solution to uncertainty modeling and online learning
- Online learning is crucial for nonstationary blind source separation

Outline

- Introduction
- Model-Based Source Separation
- Adaptive Learning Machine
- **Case Study: Independent Component Analysis**
- Case Study: Nonnegative Matrix Factorization
- Summarization and Future Trend

Case Study: Independent Component Analysis

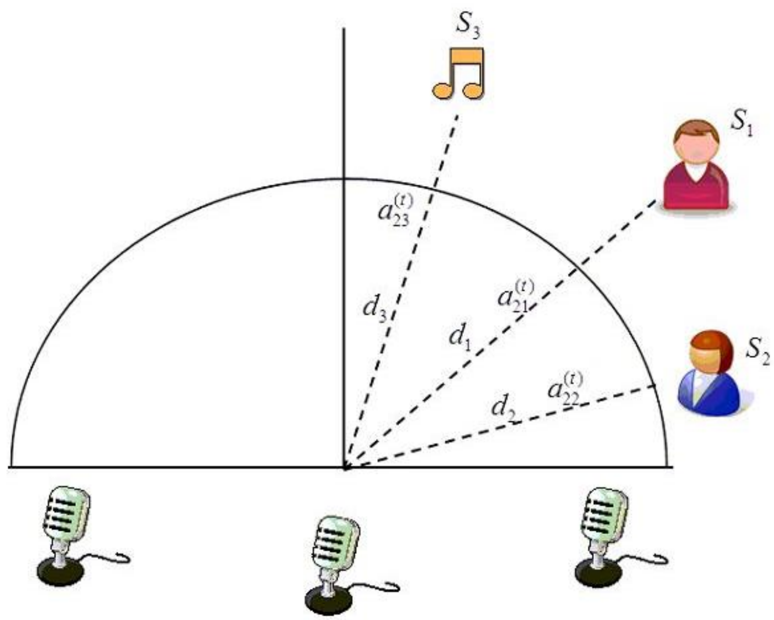
- Nonstationary Bayesian ICA
- Online Gaussian process ICA

Why nonstationary source separation?

- Real-world blind source separation
  - number of sources is unknown
  - BSS is a dynamic time-varying system
  - mixing process is nonstationary
- Why nonstationary?
  - Bayesian method using ARD can determine the changing number of sources
  - recursive Bayesian for online tracking of nonstationary conditions
  - Gaussian process provides a nonparametric solution to represent temporal structure of time-varying mixing system

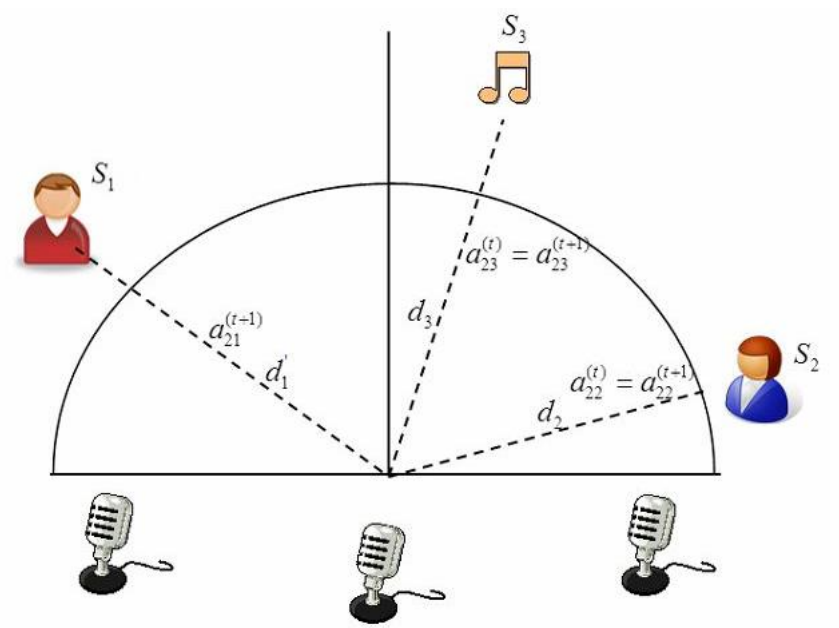
Nonstationary mixing system

- Time-varying mixing matrix is considered to reflect
  - moving sources or moving microphones
  - source signals may abruptly appear or disappear
  - source replacement



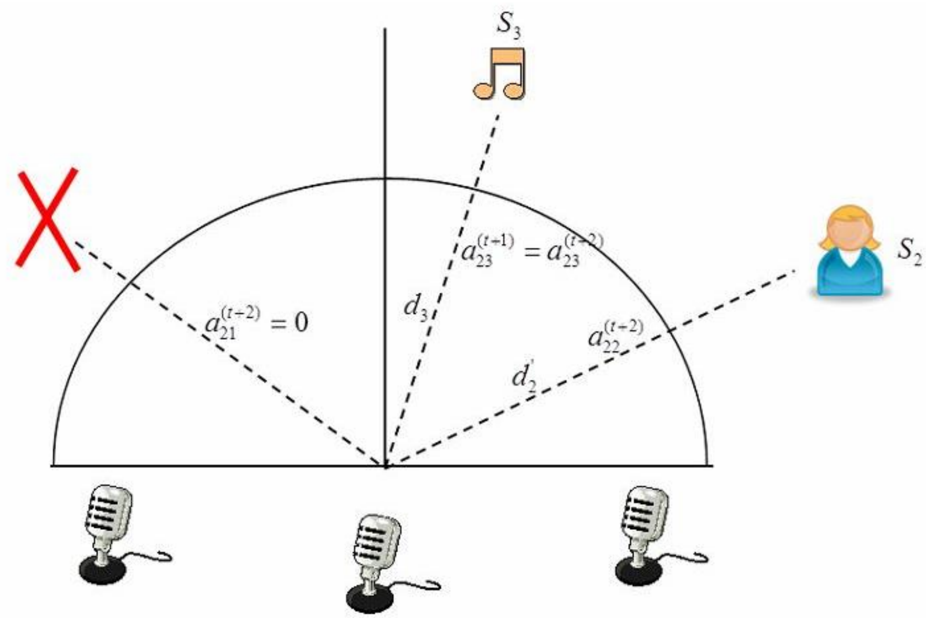
Nonstationary mixing system

- Time-varying mixing matrix is considered to reflect
  - moving sources or moving microphones
  - source signals may abruptly appear or disappear
  - source replacement



Nonstationary mixing system

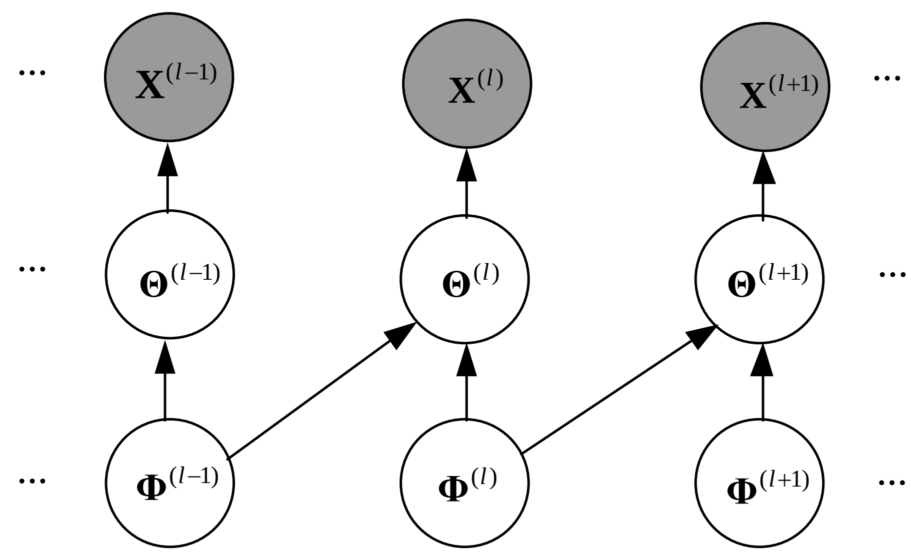
- Time-varying mixing matrix is considered to reflect
  - moving sources or moving microphones
  - source signals may abruptly appear or disappear
  - source replacement



Nonstationary Bayesian (NB) learning

- NB-ICA performs online Bayesian learning from a sequence of online minibatch training data $\mathcal{X}^{(l)} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(l)}\}$ where $\mathbf{X}^{(l)} = \{\mathbf{x}_t^{(l)}\}$

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



Model construction

- Noisy ICA model: $\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \boldsymbol{\varepsilon}_t$
- Likelihood function with time-varying mixing matrix $\mathbf{A}^{(l)}$ and source signal $\mathbf{s}^{(l)}$

$$p(\mathbf{x}_t|\mathbf{A}^{(l)}, \mathbf{s}^{(l)}, \beta^{(l)}) = \mathcal{N}(\mathbf{x}_t|\mathbf{A}^{(l)}\mathbf{s}^{(l)}, {\beta^{(l)}}^{-1} I_N)$$

- Distribution of model parameters
  - source $p(\mathbf{s}^{(l)}|\boldsymbol{\pi}^{(l)}, \boldsymbol{\mu}^{(l)}, \boldsymbol{\gamma}^{(l)}) = \prod_{m=1}^{M} \left[\sum_{k=1}^{K} \pi_k^{(l)} \mathcal{N}(s_m^{(l)}|\mu_k^{(l)}, {\gamma_k^{(l)}}^{-1})\right]$
  - mixing matrix $p(\mathbf{A}^{(l)}|\boldsymbol{\alpha}^{(l)}) = \prod_{m=1}^{M} \left[\prod_{n=1}^{N} \mathcal{N}(a_{nm}^{(l)}|0, {\alpha_m^{(l)}}^{-1})\right]$
  - noise $p(\boldsymbol{\varepsilon}_t|\beta^{(l)}) = \mathcal{N}(\boldsymbol{\varepsilon}_t|0, {\beta^{(l)}}^{-1} I_N)$

Marginal distribution

- Prior distribution
  - precision of noise $p(\beta^{(l)}|u_\beta, w_\beta) = \text{Gam}(\beta^{(l)}|u_\beta, w_\beta)$
- Marginal likelihood of NB-ICA model (Chien and Hsieh, 2013)

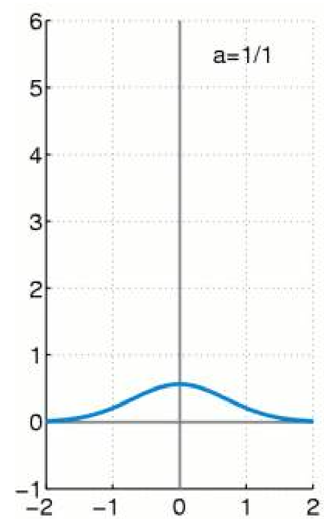
$$p(\mathbf{X}) = \prod_{t=1}^{T} \int p(\mathbf{x}_t|\mathbf{A}^{(l)}, \mathbf{s}^{(l)}, \boldsymbol{\alpha}^{(l)}, \beta^{(l)}) p(\mathbf{A}^{(l)}|\boldsymbol{\alpha}^{(l)}) p(\boldsymbol{\alpha}^{(l)}|u_\alpha^{(l)}, w_\alpha^{(l)})$$

$$\times p(\mathbf{s}^{(l)}|\boldsymbol{\pi}^{(l)}, \boldsymbol{\mu}^{(l)}, \boldsymbol{\gamma}^{(l)}) p(\beta^{(l)}|u_\beta^{(l)}, w_\beta^{(l)}) d\mathbf{A}^{(l)} d\mathbf{s}^{(l)} d\boldsymbol{\alpha}^{(l)} d\beta^{(l)}$$

Automatic relevance determination

- ARD parameter for source signals

$$\alpha_m^{(l)} = \begin{cases} \infty & , \quad a_m^{(l)} = \{a_{nm}^{(l)}\} \to 0 \\ < \infty & , \quad a_m^{(l)} = \{a_{nm}^{(l)}\} \neq 0 \end{cases}$$



  - number of sources can be determined

Compensation for nonstationary mixing

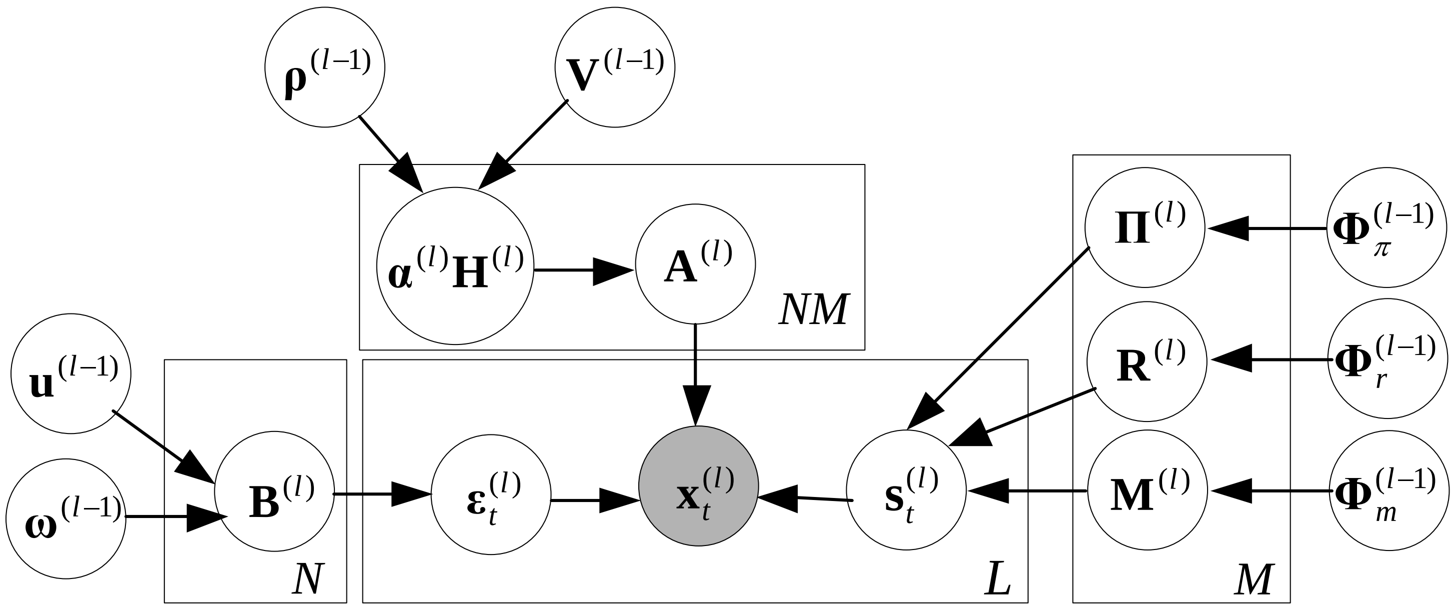
- Compensation via transformation parameter

$$G_{\mathbf{H}^{(l)}}(\boldsymbol{\alpha}^{(l)}) = \boldsymbol{\alpha}^{(l)}\mathbf{H}^{(l)}$$

- Prior for compensation parameter
  - conjugate prior using Wishart distribution

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

Graphical representation



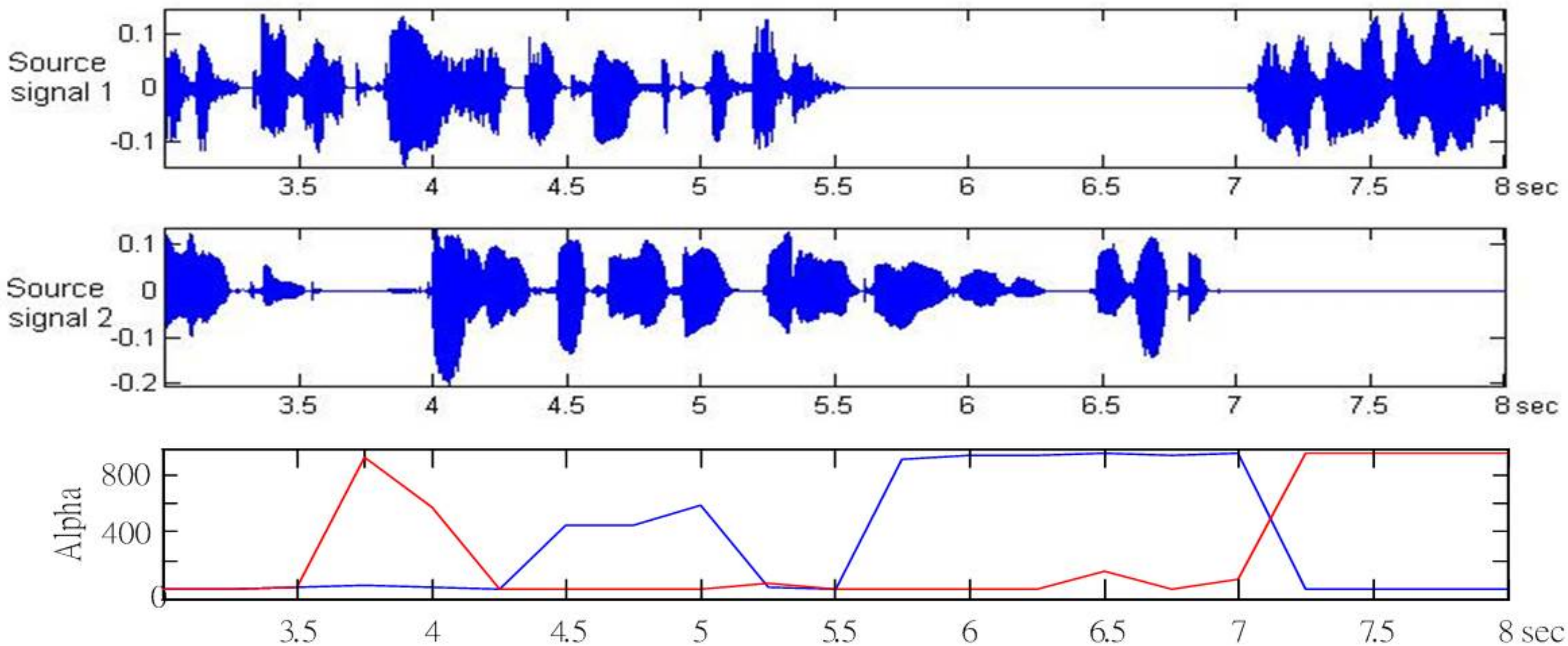
Experiment on BSS

- Experiment on nonstationary blind source separation
  - ICA'99 http://sound.media.mit.edu/ica-bench/
- Scenarios
  - state of source signals: active or inactive
  - source signals or sensors are moving: nonstationary mixing matrix

$$\mathbf{A}_t = \begin{bmatrix} \cos(2\pi f_1 t) & \sin(2\pi f_2 t) \\ -\sin(2\pi f_1 t) & \cos(2\pi f_2 t) \end{bmatrix}$$

where $f_1 = 1/5$ Hz $\quad f_2 = 1/2.5$ Hz

Source signals and ARD curves



Blue: first source signal
Red: second source signal

Case Study: Independent Component Analysis

- Nonstationary Bayesian ICA
- Online Gaussian process ICA

Online Gaussian process

- Basic ideas
  - incrementally detect the status of source signals and estimate the corresponding distributions from online observation data $\mathcal{X}^{(l)} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(l)}\}$
  - dynamic model is required to capture the temporal correlation for source separation (Smaragdis et al., 2014)
  - temporal structure of time-varying mixing coefficients $A^{(l)}$ are characterized by Gaussian process
  - Gaussian process is a nonparametric model which defines the prior distribution over functions for Bayesian inference
- Online Gaussian process (OLGP) was proposed for blind source separation (Chien and Hsieh, 2013)

Gaussian process

- GP is an infinite-dimensional generalization of multivariate normal distribution
- GP was applied to model the source signals for blind source separation (Park and Choi, 2008)
- Mixing matrix is characterized by OLGP
  - $\mathbf{A}_t^{(l)}$ is generated by a latent function $f(\cdot)$

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

$$\text{where} \quad \mathbf{a}_{nm,t-1}^{(l)} = [a_{nm,t-1}^{(l)}, \cdots, a_{nm,t-p}^{(l)}]^T$$

  - GP is adopted to describe the distribution of latent function

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

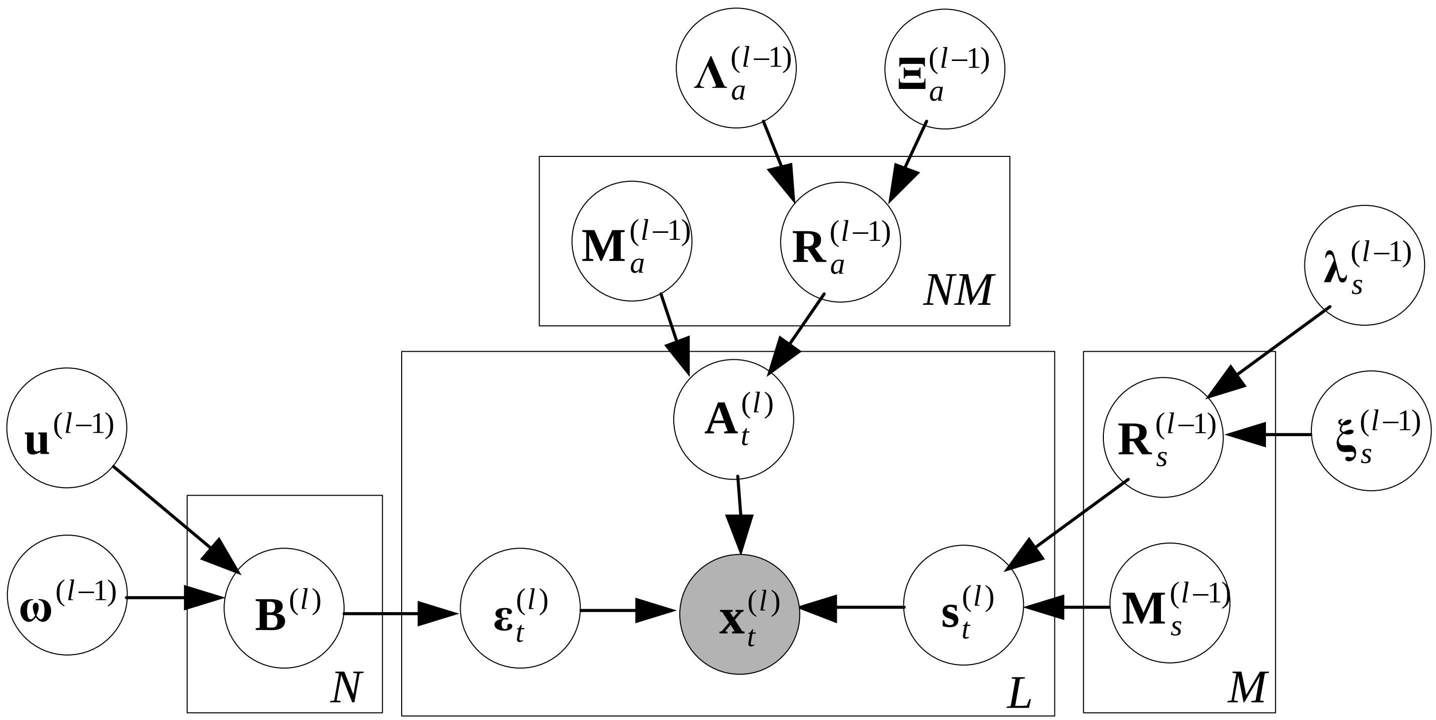
– Exponential-quadratic kernel function $\kappa(\cdot)$ is adaopted

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

– $\{\lambda_{a_{nm}}^{(l-1)}, \xi_{a_{nm}}^{(l-1)}\}$ are hyperparameters of kernel function

- GP prior could be used to represent temporal structure of time-varying source samples $\{s_{m,t}^{(l)}\}$ within a frame l
- OLGP-ICA algorithm is implemented by variational inference

Graphical representation



Experiment on BSS

- Experiment on nonstationary blind source separation
  - http://www.kecl.ntt.co.jp/icl/signal/
- Scenarios
  - state of source signals: active or inactive
  - source signals or sensors are moving: nonstationary mixing matrix

$$\mathbf{A}_t = \begin{bmatrix} \cos(2\pi f_1 t) & \sin(2\pi f_2 t) \\ -\sin(2\pi f_1 t) & \cos(2\pi f_2 t) \end{bmatrix}$$

where $f_1 = 1/20$ Hz $\quad f_2 = 1/10$ Hz

Outline

- Introduction
- Model-Based Source Separation
- Adaptive Learning Machine
- Case Study: Independent Component Analysis
- **Case Study: Nonnegative Matrix Factorization**
- Summarization and Future Trend

Case Study: Nonnegative Matrix Factorization

- Bayesian NMF
- Group sparse NMF

Why Bayesian NMF?

- Uncertainty modeling helps improving model regularization
- Uncertainties in source separation may come from
  - improper model assumption
  - incorrect model order
  - possible noise interference
  - nonstationary environment
  - reverberant distortion
  - variations of source signals
- Bayesian learning aims to build a robust source separation by maximizing the marginal likelihood over randomness of model parameters

Gaussian-Exponential BNMF

- Gaussian likelihood for modeling error (Schmidt et al., 2009)

$$p(\mathbf{X}|\mathbf{B},\mathbf{W},\sigma^2) = \prod_{m,n} \mathcal{N}(X_{mn}; [\mathbf{BW}]_{mn}, \sigma^2)$$

- Exponential prior for $\mathbf{B}$ and $\mathbf{W}$

$$p(\mathbf{B}) = \prod_{m,k} \mathsf{Exp}(B_{mk}; \lambda^b_{mk}), \;\; p(\mathbf{W}) = \prod_{k,n} \mathsf{Exp}(W_{kn}; \lambda^w_{kn})$$

- Inverse gamma prior for noise variance σ^2

$$p(\sigma^2) = \mathsf{Gam}^{-1}(\sigma^2; k, \theta)$$

Poisson-Gamma BNMF

- Poisson likelihood for $\mathbf{X}$ (Cemgil, 2009)

$$X_{mn} = \sum_k Z_{mkn}, \;\; Z_{mkn} \sim \mathsf{Pois}(Z_{mkn}; B_{mk}W_{kn})$$

$$p(\mathbf{X}|\mathbf{B}, \mathbf{W}) = \prod_{m,n} \mathsf{Pois}(X_{mn}; \sum_k B_{mk}W_{kn})$$

- Gamma prior for $\mathbf{B}$ and $\mathbf{W}$

$$p(B_{mk}; a^B_{mk}, b^B_{mk}) = \mathsf{Gam}(B_{mk}; a^B_{mk}, \frac{b^B_{mk}}{b^B_{mk}})$$

$$p(W_{kn}; a^W_{kn}, b^W_{kn}) = \mathsf{Gam}(W_{kn}; a^W_{kn}, \frac{b^W_{kn}}{a^W_{kn}})$$

Discussion

- Gibbs sampling for Gaussian-Exponential BNMF
- Variational inference for Poisson-Gamma BNMF
- Drawbacks
  - Gibbs sampling in Gaussian-Exponential BNMF and Newton's solution in Poisson-Gamma BNMF are computationally expensive
  - some dependencies during optimization were ignored
  - observations in Gaussian-Exponential BNMF are not constrained to be nonnegative

Poisson-Exponential BNMF

- Poisson likelihood for $\mathbf{X}$

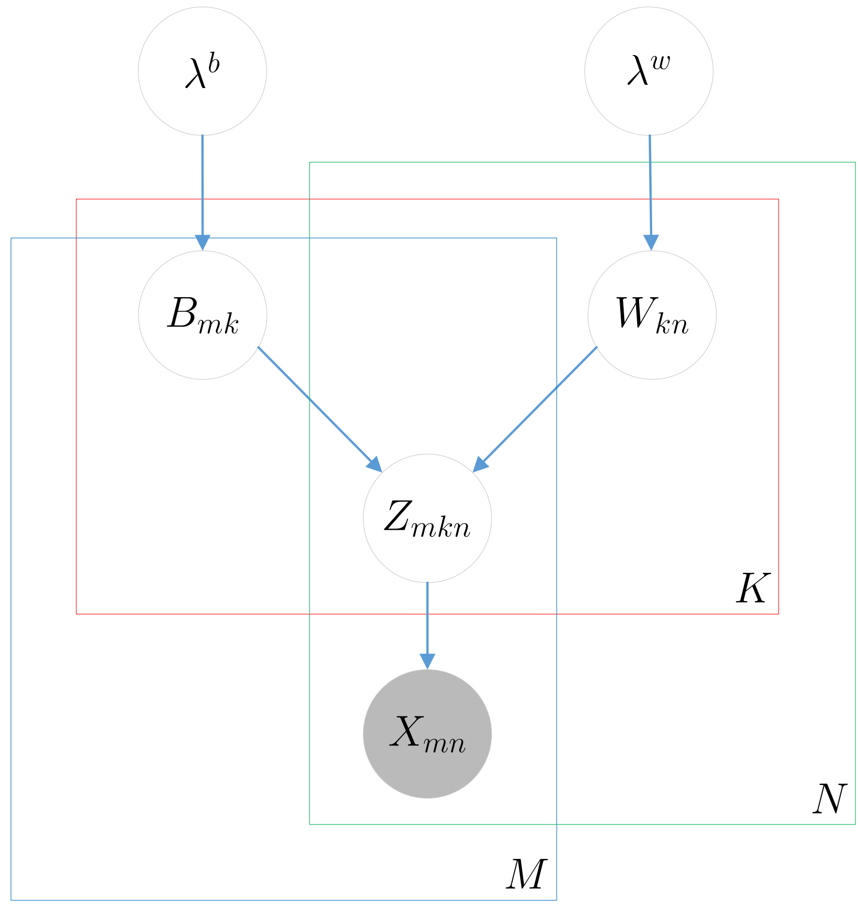
$$X_{mn} = \sum_k Z_{mkn}, \;\; p(\mathbf{X}|\mathbf{B},\mathbf{W}) = \prod_{m,n} \mathsf{Pois}(X_{mn}; \sum_k B_{mk} W_{kn})$$

- Exponential prior for $\mathbf{B}$ and $\mathbf{W}$

$$p(\mathbf{B}) = \prod_{m,k} \mathsf{Exp}(B_{mk}; \lambda^b_{mk}), \;\; p(\mathbf{W}) = \prod_{k,n} \mathsf{Exp}(W_{kn}; \lambda^w_{kn})$$

- Marginal likelihood over $\mathbf{Z}$ and $\{\mathbf{B},\mathbf{W}\}$ is optimized to find the sparsity-controlled hyperparameters $\Theta = \{\lambda^b_{mk}, \lambda^w_{mk}\}$

Graphical representation



(Yang et al., 2014)

Variational inference

- Variational distributions are derived in VB-E step as

$$q(Z_{m,:,n}) \propto \mathsf{Mult}(Z_{m,:,n}; X_{mn}, P_{m,:,n})$$
$$q(B_{mk}) \propto \mathsf{Gam}(B_{mk}; \alpha^b_{mk}, \beta^b_{mk})$$
$$q(W_{kn}) \propto \mathsf{Gam}(W_{kn}; \alpha^w_{kn}, \beta^w_{kn})$$

with variational parameters

$$\hat{\alpha}^b_{mk} = 1 + \sum_n \langle Z_{mkn} \rangle, \ \hat{\beta}^b_{mk} = \left(\sum_n \langle W_{kn} \rangle + \lambda^b_{mk} \right)^{-1}$$

$$\hat{\alpha}^w_{kn} = 1 + \sum_m \langle Z_{mkn} \rangle, \ \hat{\beta}^w_{kn} = \left(\sum_k \langle B_{mk} \rangle + \lambda^w_{kn} \right)^{-1}$$

$$\hat{P}_{mkn} = \frac{\exp(\langle \log B_{mk} \rangle + \langle \log W_{kn} \rangle)}{\sum_j \exp(\langle \log B_{mj} \rangle + \langle \log W_{jn} \rangle)}$$

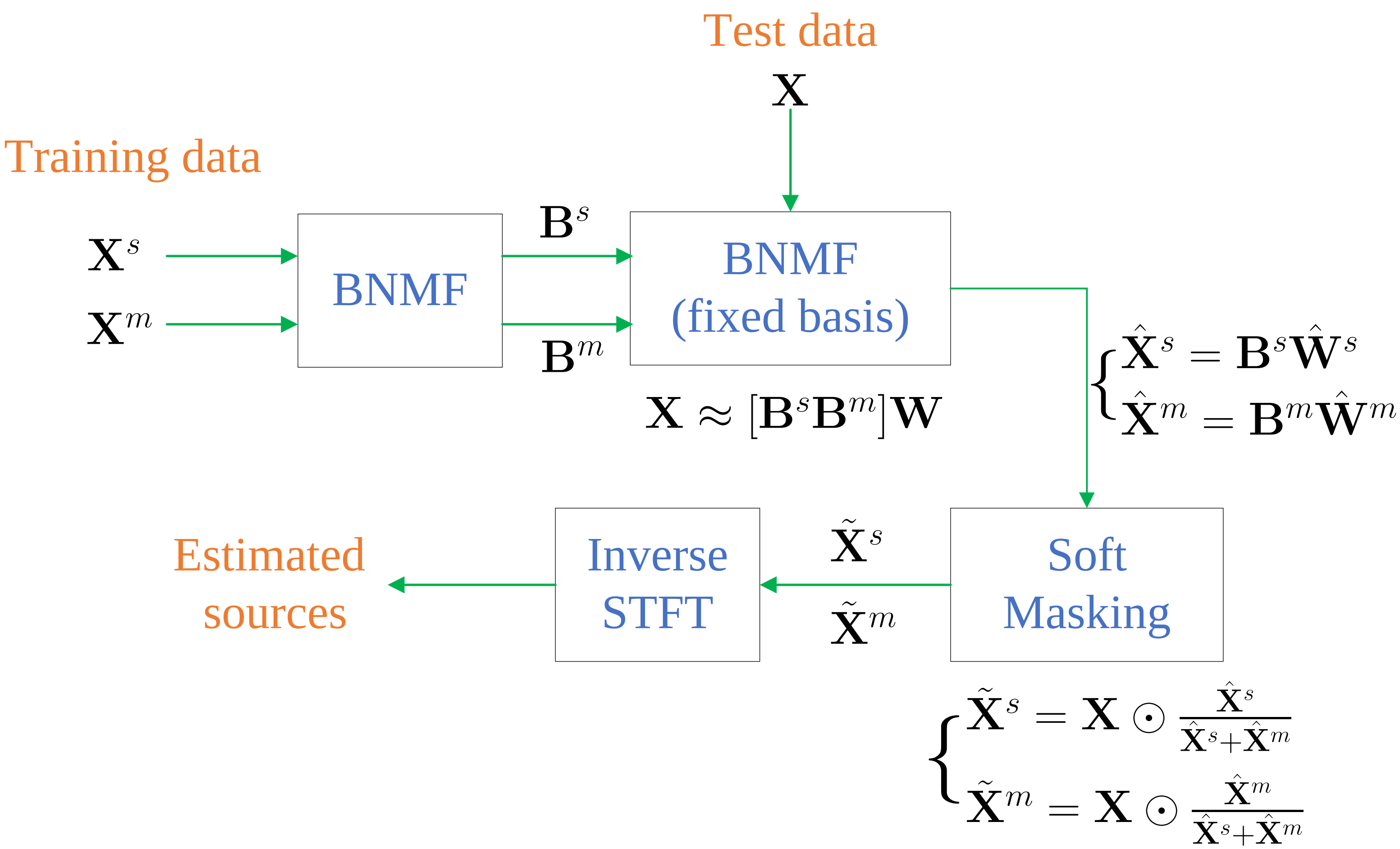
VB-M step

- Optimal regularization parameters $\Theta = \{\lambda^b_{mk}, \lambda^w_{kn}\}$ are derived by maximizing variational lower bound w.r.t. Θ

$$\hat{\lambda}^b_{mk} = \frac{1}{2}\left(-\sum_n \langle W_{kn}\rangle + \sqrt{(\sum_n \langle W_{kn}\rangle)^2 + 4\frac{\sum_n \langle W_{kn}\rangle}{\langle B_{mk}\rangle}}\right)$$

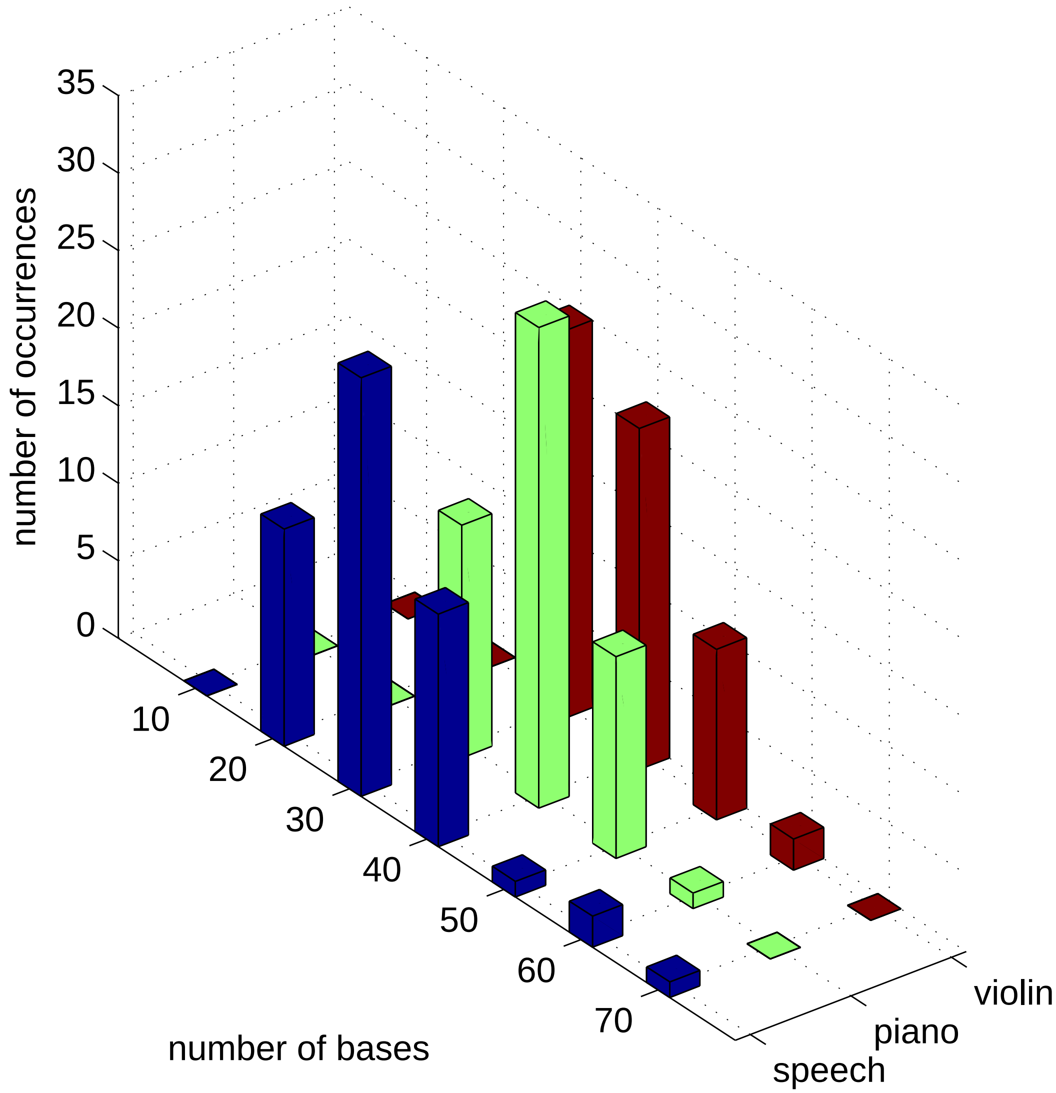
$$\hat{\lambda}^w_{kn} = \frac{1}{2}\left(-\sum_m \langle B_{mk}\rangle + \sqrt{(\sum_m \langle B_{mk}\rangle)^2 + 4\frac{\sum_m \langle B_{mk}\rangle}{\langle W_{kn}\rangle}}\right)$$

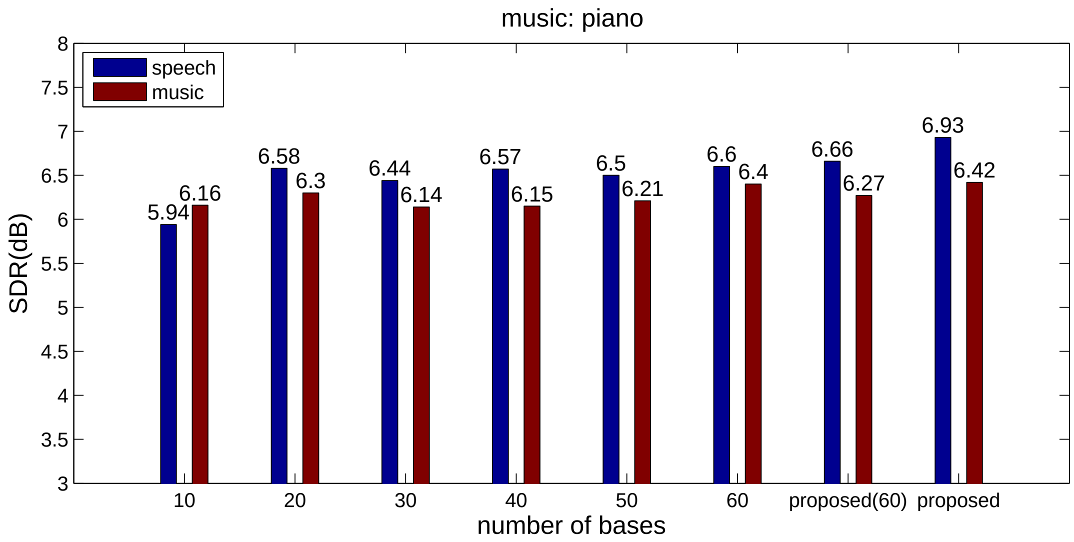
Supervised source separation

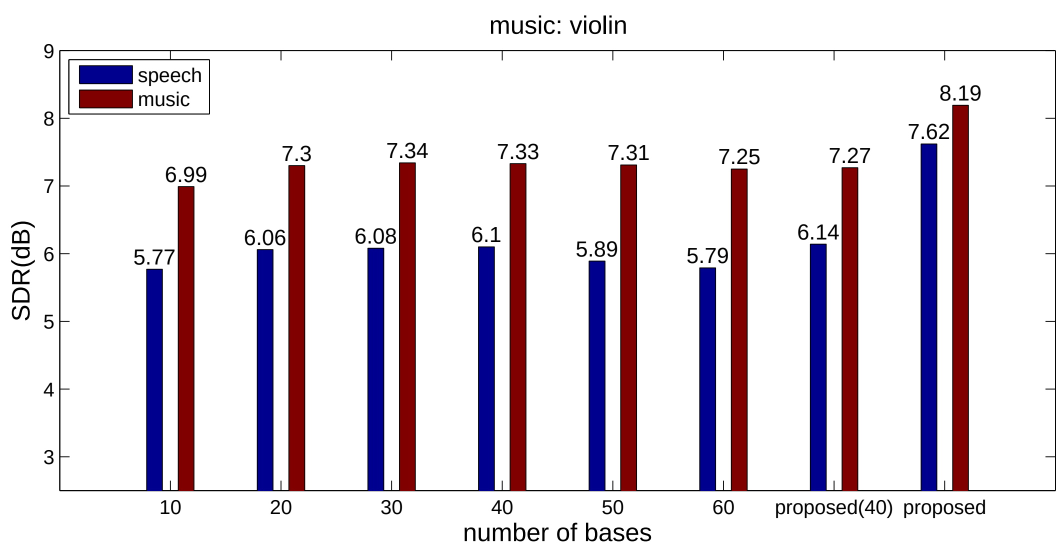


Experimental setup

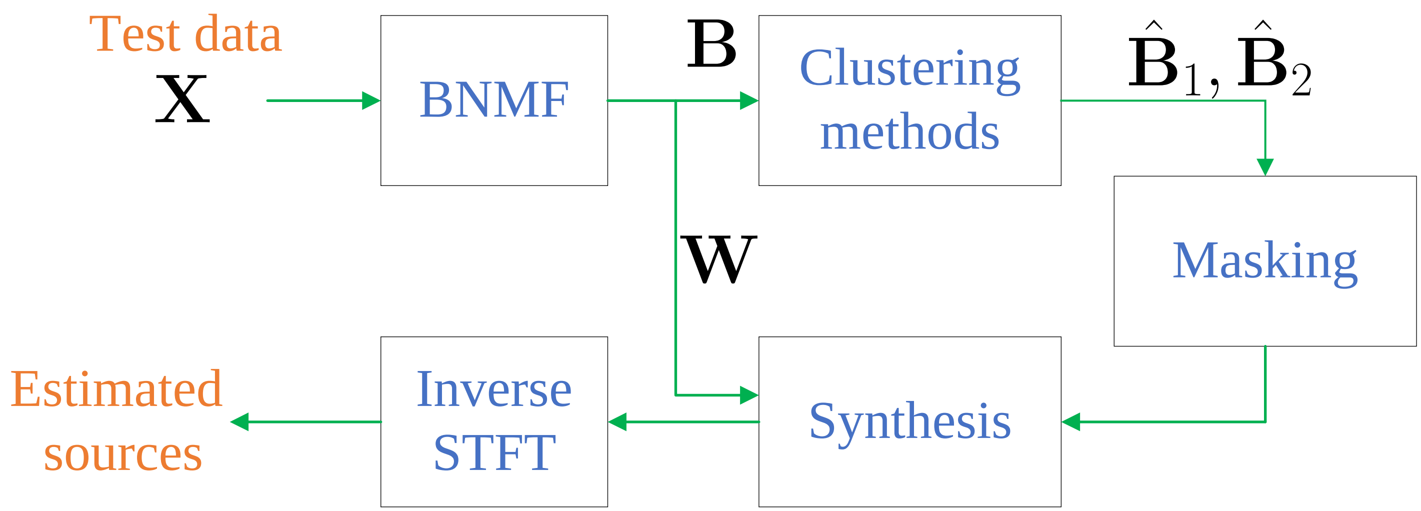
- Speech samples from TIMIT corpus
  - Randomly select 60 sentences with 3 males and 3 females
  - each sentence has a length of 2-3 seconds
- Music samples from Saarland Music Data (SMD)
  - select one piano and one violin pieces composed by Bach from the second collections
- Test signals are generated by corrupting with a randomly selected music segments at 0 dB speech-to-music ratio (SMR)
- 10-fold cross validation for each speaker
- STFT: 40ms frame duration, 10ms frame shift, 1024-points


number of occurrences
35
30
25
20
15
10
5
0
10
20
30
40
50
60
70
number of bases
speech
piano
violin


music: piano
SDR(dB)
number of bases
speech
music
10
20
30
40
50
60
proposed(60)
proposed
5.94
6.16
6.58
6.3
6.44
6.14
6.57
6.15
6.5
6.21
6.6
6.4
6.66
6.27
6.93
6.42

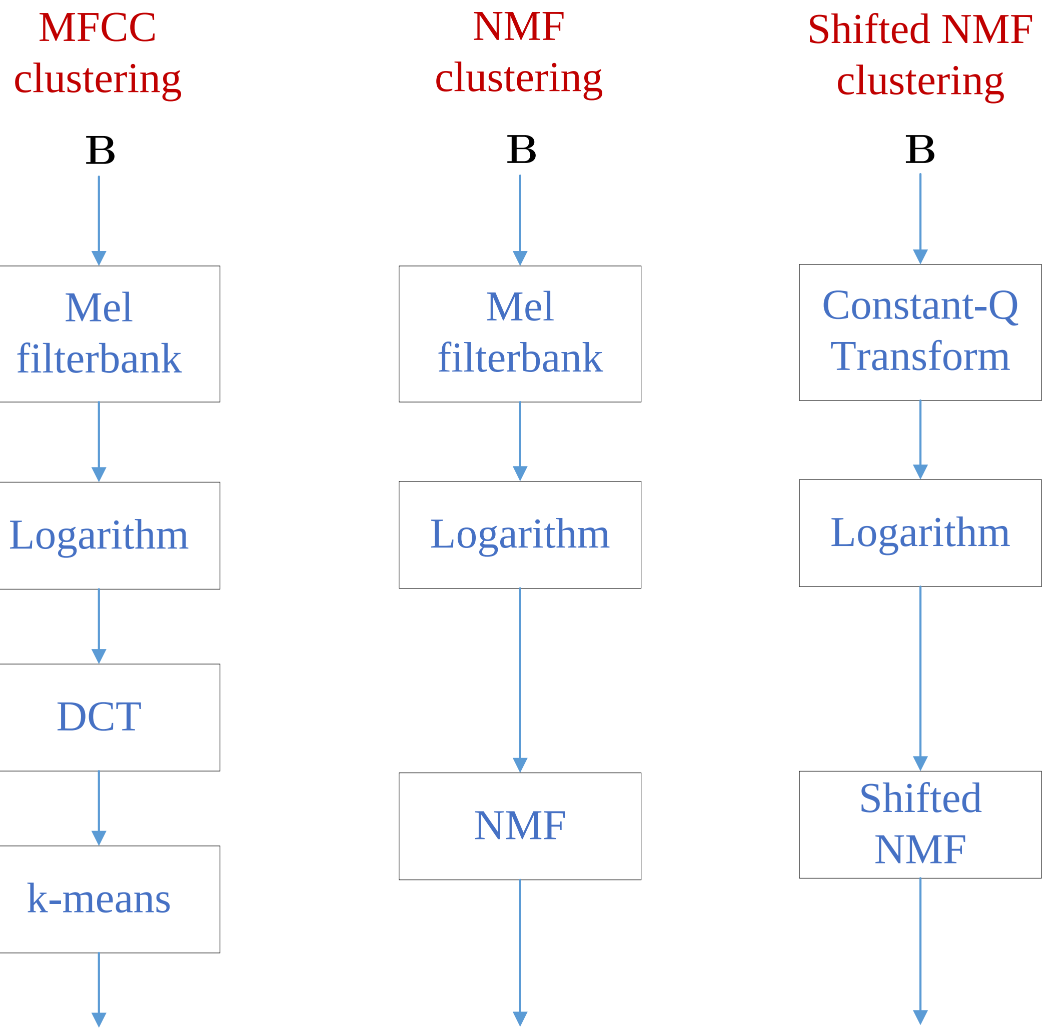

music: violin
SDR(dB)
number of bases
speech
music
10
20
30
40
50
60
proposed(40)
proposed
5.77
6.99
6.06
7.3
6.08
7.34
6.1
7.33
5.89
7.31
5.79
7.25
6.14
7.27
7.62
8.19

Unsupervised source separation



(Yang et al., 2014)

NMF clustering



- Experimental data: MIR-1K dataset
  - 1000 song clips extracted from 110 Chinese karaoke pop songs performing by 8 female and 11 male amateurs
  - Each clip recorded at 16 KHz sampling frequency with the duration ranging from 4 to 13 seconds
- SMRs of 5, 0, and -5 dB are investigated
- STFT: 40ms frame duration, 10ms frame shift, 1024-points
- Evaluation measure

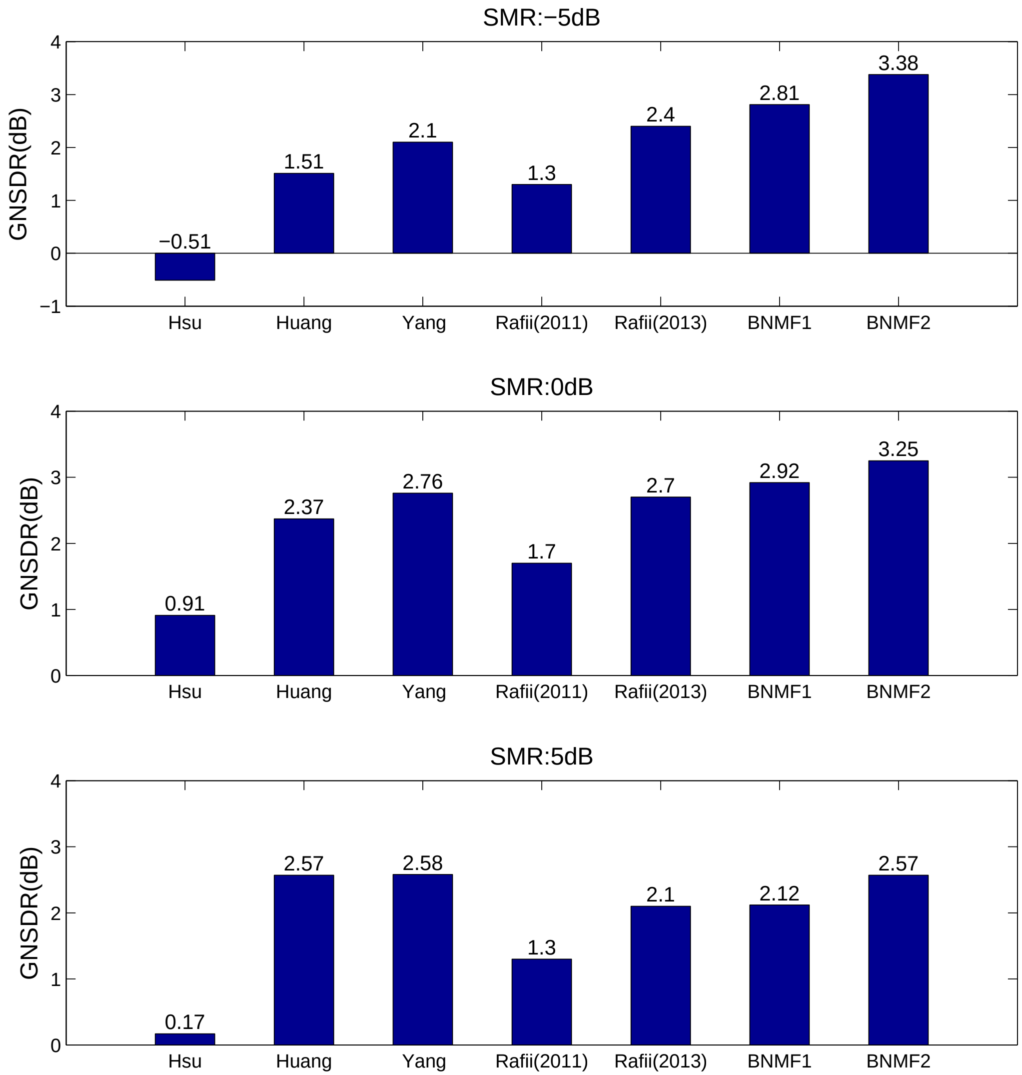
$$\text{NSDR}(\hat{\mathbf{V}}, \mathbf{V}, \mathbf{X}) = \text{SDR}(\hat{\mathbf{V}}, \mathbf{V}) - \text{SDR}(\mathbf{X}, \mathbf{V})$$

$$\text{GNSDR}(\hat{\mathbf{V}}, \mathbf{V}, \mathbf{X}) = \frac{\sum_{n=1}^{\tilde{N}} l_n \text{NSDR}(\hat{\mathbf{V}}_n, \mathbf{V}_n, \mathbf{X}_n)}{\sum_{n=1}^{\tilde{N}} l_n}$$

Evaluation

- Comparison of GNSDR at SMR $= 0$ dB using NMF with fixed number of bases $\{20, 30, 40, 50\}$ and BNMF with adaptive number of bases

	NMF (20)	NMF (30)	NMF (40)	NMF (50)	BNMF
K-means clustering	2.85	2.69	2.58	2.47	2.92
NMF clustering	3.29	3.15	3.13	2.97	3.25
Shifted NMF clustering	3.39	3.26	3.16	3.03	4.01

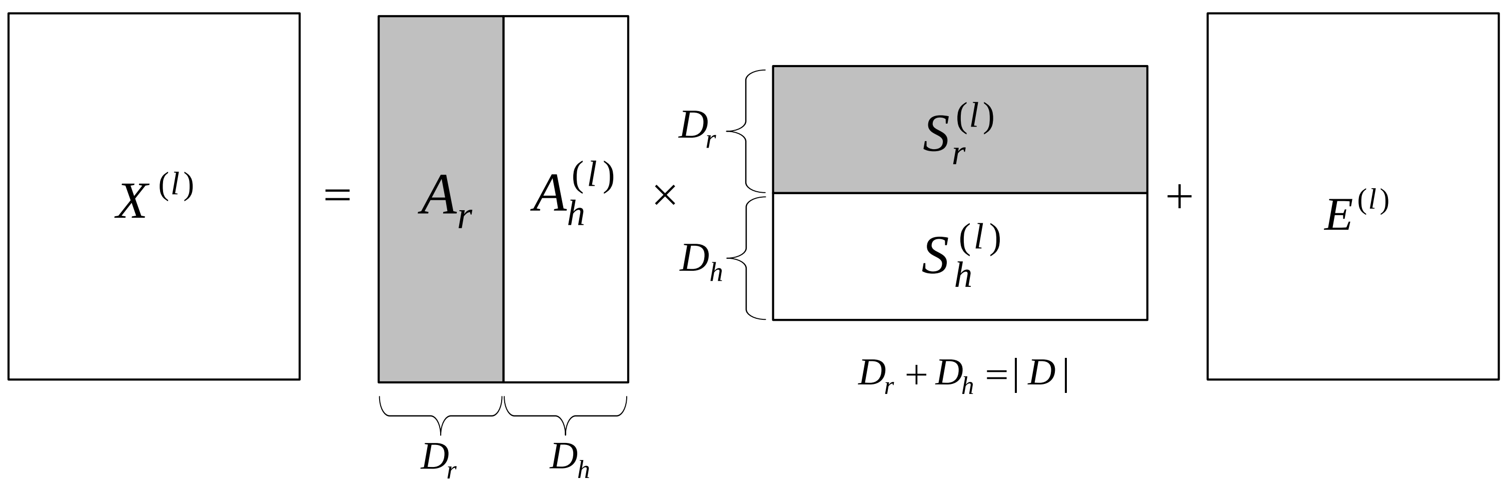

SMR:−5dB
GNSDR(dB)
−0.51
1.51
2.1
1.3
2.4
2.81
3.38
SMR:0dB
GNSDR(dB)
0.91
2.37
2.76
1.7
2.7
2.92
3.25
SMR:5dB
GNSDR(dB)
0.17
2.57
2.58
1.3
2.1
2.12
2.57
Hsu
Huang
Yang
Rafii(2011)
Rafii(2013)
BNMF1
BNMF2

Case Study: Nonnegative Matrix Factorization

- Bayesian NMF
- Group sparse NMF

Group basis representation

- Single-channel music source separation in presence of one rhythmic or repetitive signal and one harmonic or residual signal (Chien and Hsieh, 2013:18)
  - $A_r \in \mathcal{R}_+^{N \times D_r}$: shared basis matrix for all segments $\{X^{(l)}, l = 1, \ldots, L\}$
  - $A_h^{(l)} \in \mathcal{R}_+^{N \times D_h}$: individual basis matrix for segment $X^{(l)}$



$$X^{(l)} = A_r S_r^{(l)} + A_h^{(l)} S_h^{(l)} + E^{(l)}$$

Model construction

- Gaussian likelihood

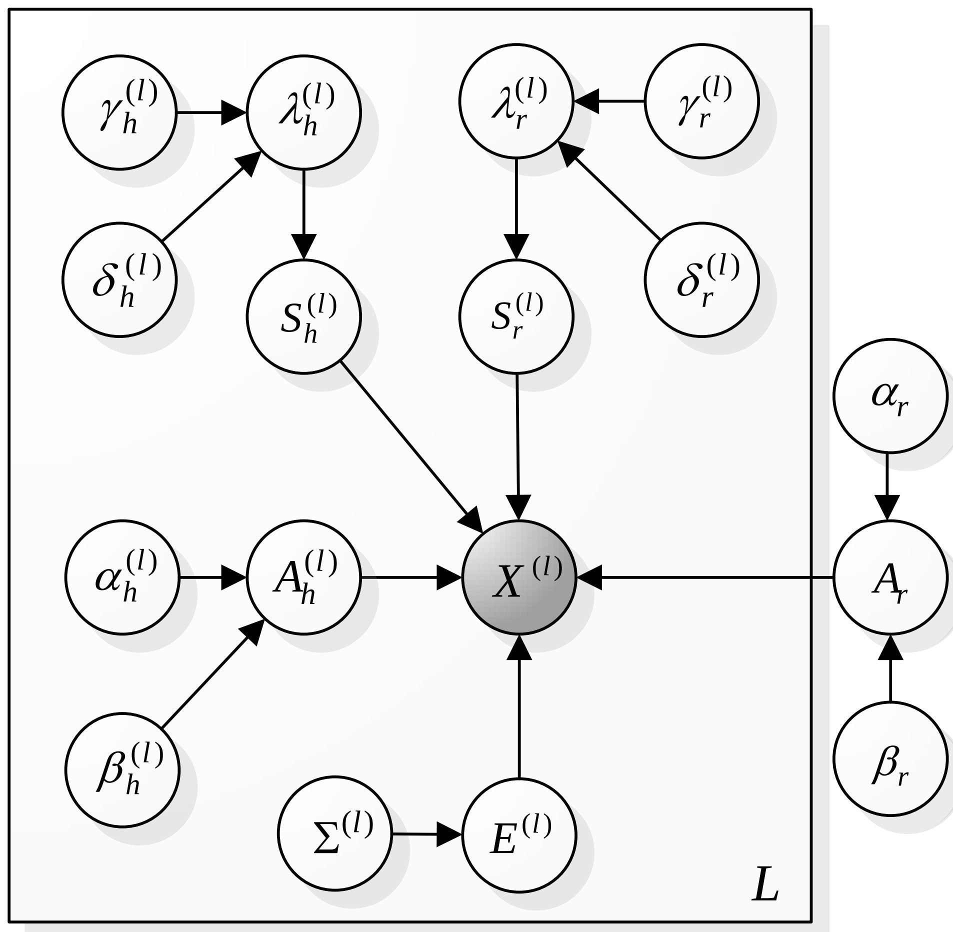
$$p(X^{(l)}|\Theta^{(l)}) = \prod_{i=1}^{N}\prod_{k=1}^{M}\mathcal{N}(X_{ik}^{(l)} \mid [A_r S_r^{(l)}]_{ik} + [A_h^{(l)} S_h^{(l)}]_{ik}, [\Sigma^{(l)}]_{ii})$$

- Gamma prior for basis parameter and Laplace prior for weight parameter

$$p(A_r) = \prod_{i=1}^{N}\prod_{j=1}^{D_r}\mathcal{G}([A_r]_{ij}|\alpha_{rj}, \beta_{rj}), \quad p(A_h^{(l)}) = \prod_{i=1}^{N}\prod_{j=1}^{D_h}\mathcal{G}([A_h^{(l)}]_{ij}|\alpha_{hj}^{(l)}, \beta_{hj}^{(l)})$$

$$p([S_r^{(l)}]_{jk}|\lambda_{rj}^{(l)}) = \frac{\lambda_{rj}^{(l)}}{2}\,\mathsf{exp}\{-\lambda_{rj}^{(l)}[S_r^{(l)}]_{jk}\}$$

Graphical representation



MCMC sampling

- MCMC sampling is developed to sequentially infer parameters $\Theta^{(t+1)}$ and hyperparameters $\Phi^{(t+1)}$ at each new iteration $t+1$ according to the posterior distribution $p(\Theta, \Phi | X)$
  - $\Theta^{(l)} = \{A_r, A_h^{(l)}, S_r^{(l)}, S_h^{(l)}, \Sigma^{(l)}\}$
  - $\Phi^{(l)} = \{\Phi_a^{(l)}, \Phi_s^{(l)}\}$
    where $\Phi_s^{(l)} = \{\gamma_{rj}^{(l)}, \delta_{rj}^{(l)}, \gamma_{hj}^{(l)}, \delta_{hj}^{(l)}\}$ and $\Phi_a^{(l)} = \{\{\alpha_{rj}, \beta_{rj}\}, \{\alpha_{hj}^{(l)}, \beta_{hj}^{(l)}\}\}$
- Nonnegativity constraint is imposed on $\{A_r, A_h^{(l)}, S_r^{(l)}, S_h^{(l)}\}$ during sampling procedure

Experiment on music source separation

- Six rhythmic signals and six harmonic signals from http://www.free-scores.com/index_uk.php3 and http://www.freesound.org/ were sampled
  - "music 1": bass+piano
  - "music 2": drum+guitar
  - "music 3": drum+violin
  - "music 4": cymbal+organ
  - "music 5": drum+saxophone
  - "music 6": cymbal+singing
- 1,000 Gibbs sampling iterations, 200 burn-in iterations
- $D_r = 15$ and $D_h = 10$

Outline

- Introduction
- Model-Based Source Separation
- Adaptive Learning Machine
- Case Study: Independent Component Analysis
- Case Study: Nonnegative Matrix Factorization
- **Summarization and Future Trend**

Summarization

- Advances in machine learning for source separation are surveyed
- Model-based blind source separation
  - independent component analysis
  - nonnegative matrix factorization
- Adaptive learning machine
  - Bayesian learning
  - sparse learning
  - online learning

Future Trend

- Source separation versus machine learning
  - DNN is powerful for BSS but in-domain signal processing is required
  - perceptual objective and measure
  - multidisciplinary approach from signal processing and machine learning
  - combined separation and classification with discriminative training

- Source separation in heterogeneous conditions
  - temporally-correlated sources
  - nonstationary mixing condition
  - adaptive model complexity
  - guided source separation, user interaction, side information (Vincent et al., 2014)

- Ubiquitous extensions and applications
  - multi-modalities, multi-models and multi-ways in source separation

Thanks to

Chung-Chien Hsu
Po-Kai Yang
Guan-Xiang Wang
Hsin-Lung Hsieh

Machine Learning Lab, National Chiao Tung University

& Ministry of Science and Technology, Taiwan

References

[1] S. Amari, “Natural gradient works efficiently in learning,” *Neural Computation*, vol. 10, pp. 251–276, 1998.

[2] S. Araki, T. Nakatani, H. Sawada, and S. Makino, “Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior,” in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 33–36.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer, 2008.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

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

[6] J.-T. Chien and H.-L. Hsieh, “Bayesian group sparse learning for music source separation,” *EURASIP Journal on Audio, Speech, and Music Processing*, 2013:18.

[7] ——, “Nonstationary source separation using sequential and variational Bayesian learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 681–694, 2013.

[8] P. Comon, “Independent component analysis, a new concept?” *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[9] C. Fevotte, “Bayesian audio source separation,” in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer, 2007, pp. 305–335.

[10] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[11] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[12] Y. Li, Z. L. Yu, N. Bi, Y. Xu, Z. Gu, and S. Amari, “Sparse representation for brain signal processing - a tutorial on methods and applications,” *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 96–106, 2014.

[13] S. Park and S. Choi, "Gaussian processes for source separation," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 1909–1912.

[14] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition - graphical modeling approaches," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 66–80, 2010.

[15] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer, 2007, pp. 47–78.

[16] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," in *Proc. of International Conference on Independent Component Analysis and Signal Separation*, 2009, pp. 540–547.

[17] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorization - a unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[18] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[19] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[20] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation - how models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[21] P.-K. Yang, C.-C. Hsu, and J.-T. Chien, "Bayesian singing-voice separation," in *Proc. of Annual Conference of International Society for Music Information Retrieval (ISMIR)*, 2014, pp. 507–512.

[22] ——, "Bayesian factorization and selection for speech and music separation," in *Proc. of Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014, pp. 998–1002.

[23] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms - robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.