2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

INFINITE PROBABILISTIC LATENT COMPONENT ANALYSIS FOR AUDIO SOURCE SEPARATION

Kazuyoshi Yoshii[1,2] *Eita Nakamura*[1] *Katsutoshi Itoyama*[1] *Masataka Goto*[3]

[1]Kyoto University [2]RIKEN [3]National Institute of Advanced Industrial Science and Technology (AIST)
{yoshii, enakamura, itoyama}@sap.ist.i.kyoto-u.ac.jp m.goto@aist.go.jp

ABSTRACT

This paper presents a statistical method of audio source separation based on a nonparametric Bayesian extension of probabilistic latent component analysis (PLCA). A major approach to audio source separation is to use nonnegative matrix factorization (NMF) that approximates the magnitude spectrum of a mixture signal at each frame as the weighted sum of fewer source spectra. Another approach is to use PLCA that regards the magnitude spectrogram as a two-dimensional histogram of "sound quanta" and classifies each quantum into one of sources. While NMF has a physically-natural interpretation, PLCA has been used successfully for music signal analysis. To enable PLCA to estimate the number of sources, we propose Dirichlet process PLCA (DP-PLCA) and derive two kinds of learning methods based on variational Bayes and collapsed Gibbs sampling. Unlike existing learning methods for nonparametric Bayesian NMF based on the beta or gamma processes (BP-NMF and GaP-NMF), our sampling method can efficiently search for the optimal number of sources without truncating the number of sources to be considered. Experimental results showed that DP-PLCA is superior to GaP-NMF in terms of source number estimation.

Index Terms— Source separation, nonparametric Bayes, probabilistic latent component analysis, Dirichlet process, Gibbs sampling, variational Bayes.

1. INTRODUCTION

Statistical matrix factorization forms the basis of modern signal processing. Given a matrix $\boldsymbol{X} \in \mathbb{R}^{M \times N}$, the typical goal of factorization is to estimate two matrices $\boldsymbol{A} \in \mathbb{R}^{M \times K}$ and $\boldsymbol{B} \in \mathbb{R}^{K \times N}$ such that $\boldsymbol{X} \approx \boldsymbol{AB}$, where $K \ll M, N$. Such low-rank representation with a limited degree of freedom enables us to extract essential information from the original redundant data $\boldsymbol{X}$. To perform matrix factorization, it is important to carefully examine the statistical characteristics of $\boldsymbol{X}$, $\boldsymbol{A}$, and $\boldsymbol{B}$ and define an appropriate likelihood function of $\boldsymbol{A}$ and $\boldsymbol{B}$ for $\boldsymbol{X}$ to be maximized. In principal component analysis (PCA), for example, $\boldsymbol{X}$ is a set of observed variables, $\boldsymbol{B}$ is a set of the corresponding latent variables that are Gaussian distributed, $\boldsymbol{A}$ is an orthogonal transformation matrix, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are obtained by maximizing the Gaussian likelihood for $\boldsymbol{X}$. In independent component analysis (ICA) [1], $\boldsymbol{X}$ is a set of mixture signals, $\boldsymbol{B}$ is a set of source signals that are independently super-Gaussian (*e.g.*, Laplacian) distributed, $\boldsymbol{A}$ is a mixing matrix, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are obtained by maximizing the Gaussian likelihood for $\boldsymbol{X}$.

In the field of music signal processing, nonnegative matrix factorization (NMF) [2] and probabilistic latent component analysis (PLCA) [3], which restrict $\boldsymbol{X}$, $\boldsymbol{A}$, and $\boldsymbol{B}$ to nonnegative matrices, have been widely used for source separation of music audio signals. Note that PLCA has often been mistakenly understood as a probabilistic version of NMF. More precisely and technically, the probabilistic models underlying NMF and PLCA are a *factor model* and a *mixture model*, respectively. This makes a clear difference in how to model the observed data $\boldsymbol{X}$ (Fig. 1). A factor model represents each "sample" in $\boldsymbol{X}$ as a weighted sum of all sources and can be used for solving a decomposition problem. A mixture model, on the other hand, assumes each "sample" to exclusively belong to one of the sources and can be used for solving a clustering problem.

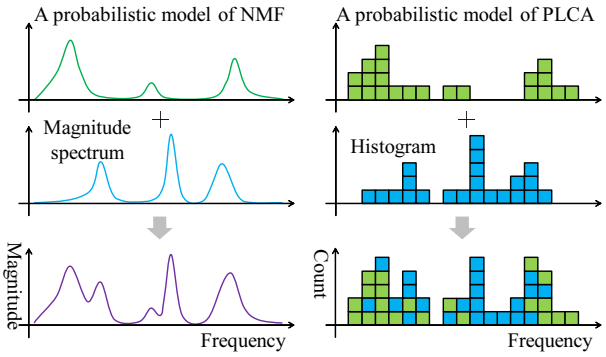

Fig. 1. A comparison of NMF and PLCA. NMF is a factor model based on the *sum of random variables* (values of magnitude) generated from probability distributions (*e.g.,* Poisson distributions) and PLCA is a mixture model based on the *sum of probability distributions* (*i.e.*, categorical distribution) used for generating random variables (sound quanta).

When NMF is used for source separation, the magnitude spectrum of each frame in an observed mixture spectrogram is regarded as a "sample" and is approximated as the sum of source spectra. Various probabilistic models of NMF can be formulated by specifying a probability distribution that generates the sample. Among others, Euclidean NMF (EU-NMF) based on the Gaussian distribution, Kullback-Leibler NMF (KL-NMF) based on the Poisson distribution [4], and Itakura-Saito NMF (IS-NMF) based on the complex Gaussian distribution [5] are popular variants of NMF. Although IS-NMF can be theoretically justified, KL-NMF has been empirically known to work best for source separation. In KL-NMF, the magnitude is assumed to be Poisson distributed and must take an integer value (can take real value in practice). This implies a deep connection between KL-NMF and PLCA.

When PLCA is used for source separation, the magnitude spectrogram in the time-frequency plane is regarded as a two-dimensional

This work was partly supported by JST ACCEL No. JPMJAC1602 and JSPS KAKENHI Numbers 26700020 and 16H01744, Japan.

978-1-5090-6341-3/17/$31.00 ©2017 IEEE

histogram of "sound quanta," each of which is assumed to be a sample generated from one of the sources. If each time-frequency bin is regarded as a single sample as in NMF, it cannot belong to multiple sources. Instead, each bin is considered to include multiple sources by collecting the sound quanta generated from multiple sources. This idea was inspired by topic models proposed for natural language processing. While each word in a document is generated from one of the topics, the document can be considered to include multiple topics. PLCA was named after a basic topic model called probabilistic latent semantic analysis (PLSA) [6].

Considering the difference between NMF and PLCA, we propose nonparametric Bayesian PLCA based on the Dirichlet process (DP) called DP-PLCA that can estimate the number of sources. The beta or gamma process (BP or GaP), on the other hand, has been used for formulating nonparametric Bayesian factor models such as BP-NMF [7, 8] and GaP-NMF [9]. While in theory infinitely many sources are supposed to exist, only a limited number of them are effectively used for representing the finite observed data. We derive two kinds of learning methods for DP-PLCA. One is a deterministic method based on variational Bayes and the other is a stochastic method based on collapsed Gibbs sampling. Unlike existing learning methods for BP-NMF and GaP-NMF, our sampling method can efficiently search for the optimal number of sources without truncating the number of sources to be considered. In this paper, we examine the capability of DP-PLCA for estimating the number of sources.

2. PRIOR ART

This section reviews nonparametric Bayesian infinite mixture and factor models of matrix factorization based on the Dirichlet, gamma, and beta processes (DP, GaP, and BP).

2.1. Infinite mixture models

An infinite mixture model (*e.g.*, Gaussian mixture model) for N observations $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ is defined by an infinite number of component distributions (*e.g.*, Gaussian distribution) $\{p(\boldsymbol{x}|\boldsymbol{\phi}_k)\}_{k=1}^\infty$ with mixing ratios $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^\infty$ as follows:

$$p(\boldsymbol{x}_n|\boldsymbol{\Theta}) = \sum_{k=1}^{\infty} \pi_k p(\boldsymbol{x}_n|\boldsymbol{\phi}_k), \tag{1}$$

where $\boldsymbol{\phi} = \{\boldsymbol{\phi}_k\}_{k=1}^\infty$ is a set of component parameters (*e.g.*, mean and covariance matrix) and $\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \boldsymbol{\phi}\}$ is a set of all parameters. Using a Dirichlet process (DP) prior $\mathrm{DP}(G_0, \alpha)$ with a base measure G_0 (*e.g.*, Gaussian-Wishart distribution) and a concentration parameter α, the generative process of $\boldsymbol{x}_n$ is represented as follows:

$$G \sim \mathrm{DP}(G_0, \alpha), \tag{2}$$

$$\hat{\boldsymbol{\phi}}_n \sim G, \quad \boldsymbol{x}_n \sim p(\boldsymbol{x}_n|\hat{\boldsymbol{\phi}}_n), \tag{3}$$

where $\hat{\boldsymbol{\phi}}_n \in \boldsymbol{\phi}$ is the parameter of a component distribution used for generating $\boldsymbol{x}_n$ and G can be explicitly written as the sum of infinitely many Dirac measures as follows (Fig. 2):

$$\boldsymbol{\phi}_k \sim G_0, \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\phi}_k}, \tag{4}$$

2.1.1. Stick-breaking process

A popular constructive interpretations of the DP is a stick-breaking process (SBP) [10] that generates the mixing ratios $\boldsymbol{\pi}$ by recursively breaking off a stick with a unit length such that the lengths of the fragments, $\boldsymbol{\pi}$, sum to unity. This process is governed by the concentration parameter α as follows (Fig. 3):

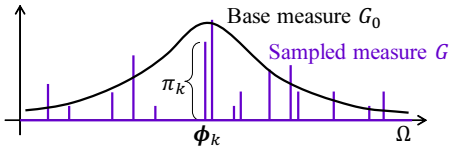

Fig. 2. A probability measure G drawn from a DP with a base measure G_0 and a concentration parameter α.

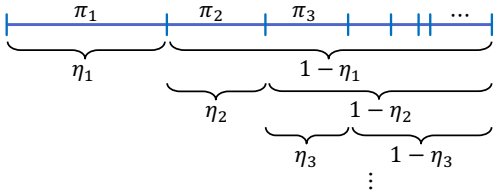

Fig. 3. A stick-breaking process for recursively generating infinitely many weights $\boldsymbol{\pi}$ that sum to unity.

$$\eta_k \sim \mathrm{Beta}(1, \alpha), \quad \pi_k = \eta_k \prod_{k'=1}^{k-1} (1 - \eta_{k'}). \tag{5}$$

More simply, the SBP is often written as follows:

$$\boldsymbol{\pi} \sim \mathrm{SBP}(\alpha). \tag{6}$$

Using the SBP, Blei and Jordan [11] proposed a deterministic learning method based on variational Bayes (VB) that approximates the complicated true posterior of the parameters $\boldsymbol{\Theta}$ as a tractable factorized distribution that can be iteratively optimized. Another approximation required in practice is that the number of components considered should be truncated at a sufficiently large level because infinitely many parameters cannot be dealt with in reality. The method is thus initialized with sufficiently many components and unnecessary components are gradually removed in each iteration to estimate the appropriate number of components, K^+

2.1.2. Chinese restaurant process

Another constructive interpretation of the DP is known as a Chinese restaurant process (CRP) [12] that sequentially generates component parameters $\hat{\boldsymbol{\phi}} = \{\hat{\boldsymbol{\phi}}_n\}_{n=1}^N$ ($\hat{\boldsymbol{\phi}}_n \in \boldsymbol{\phi}$) used for generating $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ as follows:

$$\begin{aligned}\hat{\boldsymbol{\phi}}_{n+1}|\hat{\boldsymbol{\phi}}_1, \cdots, \hat{\boldsymbol{\phi}}_n &\sim \frac{1}{\alpha+n} \sum_{n'=1}^{n} \delta_{\hat{\boldsymbol{\phi}}_{n'}} + \frac{\alpha}{\alpha+n} G_0 \\ &= \frac{1}{\alpha+n} \sum_{k=1}^{K_n} n_k \delta_{\boldsymbol{\phi}_k} + \frac{\alpha}{\alpha+n} G_0, \end{aligned} \tag{7}$$

where K_n is the number of different components (classes) used for generating $\{\boldsymbol{x}_{n'}\}_{n'=1}^n$ and n_k is the number of samples generated from class k. More simply, we often say

$$\hat{\boldsymbol{\phi}} \sim \mathrm{CRP}(G_0, \alpha) \quad \text{or} \quad \boldsymbol{Z} \sim \mathrm{CRP}(\alpha), \tag{8}$$

where $\boldsymbol{Z} = \{\boldsymbol{z}_n\}_{n=1}^N$ is a set of latent variables (class indicators) and each $\boldsymbol{z}_n$ is represented as a one-hot vector. If $\boldsymbol{x}_n$ is generated from component k, *i.e.*, $\hat{\boldsymbol{\phi}}_n = \boldsymbol{\phi}_k$, the k-th dimension of $\boldsymbol{z}_n$ takes 1 and the other dimentions take 0.

A key advantage of this representation involving only $\hat{\phi}$ is that it is unnecessary to deal with the infinite-dimensional vector $\boldsymbol{\pi}$ by marginalizing out G from Eq. (2) and Eq. (3). Instead, we consider at most a finite number of components $\hat{\phi}$ actually used for generating a finite amount of observed data $\boldsymbol{X}$.

Using the CRP, Neal [13] proposed a stochastic learning method based on Gibbs sampling (GS) that is used for generating samples (values of $\boldsymbol{\Theta}$) from the complicated posterior of $\boldsymbol{\Theta}$ without calculating its intractable normalizing factor. Unlike the VB method, the effective number of component K^+ used for representing $\boldsymbol{X}$ can be stochastically estimated in each iteration and finally the posterior of K is obtained. Although the convergence is often hard to judge, in general the GS method is more efficient (more iterations are needed, but each iteration can be performed much faster) and more robust to local maxima than the VB method.

2.1.3. *Topic models*

Topic models are an important family of mixture models originally used for natural language processing. The most basic model is probabilistic latent semantic analysis (PLSA) [6]. Let M be the number of different words in a dictionary. Given a set of N documents as observed data $\boldsymbol{X}$, PLSA aims to estimate K topics (*i.e.*, unigram probabilities of M words), $\{\{p(m|k)\}_{m=1}^{M}\}_{k=1}^{K}$, and the mixing ratios of those topics $\{\{p(k|n)\}_{k=1}^{K}\}_{n=1}^{N}$ in an unsupervised manner. The probabilistic model of PLSA is often written as follows:

$$p(n,m) = \sum_{k=1}^{K} p(m|k)p(k|n)p(n) \tag{9}$$

$$= \sum_{k=1}^{K} p(m|k)p(n|k)p(k), \tag{10}$$

where $p(n,m)$ is a joint probability distribution used for generating $\boldsymbol{X}$, *i.e.*, a two-dimensional histogram of word counts over N documents and M words and $p(n)$ is a probability distribution over N documents, which is set to the empirical ratios of word counts over N documents. Eq. (9) is usually converted to an equivalent representation given by Eq. (10) for mathematical convenience. The parameters, $\boldsymbol{\Theta}$, of all the three categorical distributions $p(m|k)$, $p(n|k)$, and $p(k)$ can be estimated by using the expectation-maximization (EM) algorithm such that the likelihood of $\boldsymbol{\Theta}$ for $\boldsymbol{X}$ given by Eq. (10) is maximized (maximum likelihood estimation).

To deal with new documents that are not included in $\boldsymbol{X}$, *i.e.*, to formulate a complete generative model of documents, a Bayesian extension of PLSA called latent Dirichlet allocation (LDA) was proposed [14]. LDA is based on Eq. (9) and assumes that $p(n)$ is a uniform distribution. Putting Dirichlet priors on $p(m|k)$ and $p(k|n)$, respectively, a topic distribution $p(k|n')$ for a new document n' can be generated.

To estimate the effective number of topics K^+, Teh *et al.* [15] proposed nonparametric Bayesian LDA that puts a hierarchical DP (HDP) prior on a set of infinitely many topic distributions (mixing ratios) $\{\{p(k|n)\}_{k=1}^{\infty}\}_{n=1}^{N}$ and word distributions (component distributions) $\{\{p(m|k)\}_{m=1}^{M}\}_{k=1}^{\infty}$. If an independent DP prior is put for each document n, a document-specific topic distribution $\{p(k|n)\}_{k=1}^{\infty}$ and word distributions $\{\{p(m|k,n)\}_{m=1}^{M}\}_{k=1}^{\infty}$ are generated. To avoid this, a higher-level DP prior is put on document-wise DP priors for sharing $\{\{p(m|k)\}_{m=1}^{M}\}_{k=1}^{\infty}$ over all documents.

2.1.4. *Probabilistic latent component analysis*

Probabilistic latent component analysis (PLCA) [3] is a generalization of PLSA for D-dimensional data ($D \geq 1$). The probabilistic model of PLCA is given by

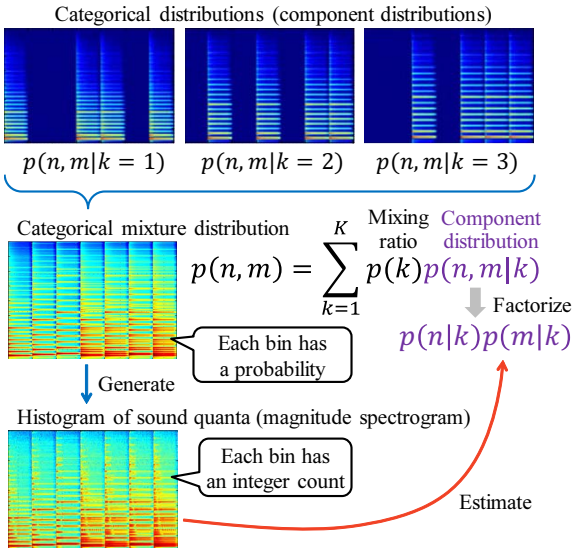

Fig. 4. The overview of PLCA. The magnitude spectrogram is assumed to follow a two-dimensional categorical mixture distribution.

$$p(x_1,\cdots,x_D) = \sum_{k=1}^{K} p(k)p(x_1,\cdots,x_D|k)$$

$$= \sum_{k=1}^{K} p(k) \prod_{d=1}^{D} p(x_d|k), \tag{11}$$

where x_d is a random variable of *any* type in dimension d ($1 \leq d \leq D$) and a sample generated from Eq. (11) is represented by a tuple $(x_1,\cdots,x_D)$. When $d = 2$ and x_1 and x_2 are discrete random variables, PLCA reduces to PLSA (x_1 and x_2 correspond to n and m, respectively, as shown in Fig. 4).

PLCA have successfully been used for audio source separation by regarding n and m as time and frequency indices, respectively [3, 16, 17]. A pair (n,m) denotes the time-frequency position of a "sound quantum" generated from Eq. (11) or Eq. (10). Given the magnitude spectrogram of a music signal, *i.e.*, a two-dimensional histogram of sound quanta, in the time-frequency plane with N frames and M frequency bins, the frequency distributions of K sources, $\{\{p(m|k)\}_{m=1}^{M}\}_{k=1}^{K}$, and their mixing ratios, $\{\{p(k|n)\}_{k=1}^{K}\}_{n=1}^{N}$, can be estimated by the EM algorithm as in PLSA. The frequency distributions can be trained from isolated musical instrument sounds in advance [16, 17]. Although HDP-PLCA could be formulated as in HDP-LDA, in this paper we show that a simpler and more easy-to-implement extension called DP-PLCA is feasible by using Eq. (11) (Eq. (10) instead of Eq. (9)) and a DP prior for efficient Bayesian inference based on VB or GS.

2.2. Infinite factor models

A finite factor model (*e.g.*, KL-NMF) for N observations $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$ is defined by a particular probability distribution (*e.g.*, Poisson distribution) that is specified by the sum of basis parameters (factors) $\boldsymbol{\phi} = \{\boldsymbol{\phi}_k\}_{k=1}^{K}$ with their *local* weights $\boldsymbol{\omega}_n = \{\omega_{nk}\}_{k=1}^{K}$ specific to sample n as follows:

$$p(\boldsymbol{x}_n|\boldsymbol{\Theta}) = p\left(\boldsymbol{x}_n \middle| \sum_{k=1}^{K} \omega_{nk}\boldsymbol{\phi}_k\right). \tag{12}$$

The gamma process (GaP) [9] or beta process (BP) [7, 8] can be used for formulating an infinite factor model by taking the limit of Eq. (12) when K goes to infinity. A key feature of infinite factor models is that although in theory infinitely many factors are assumed to exist, at most a finite number of factors are effectively used for representing $\boldsymbol{X}$. The BP was also used for estimating the number of basis patterns (atoms) in dictionary learning from images [18, 19].

2.2.1. Gamma process

Introducing a K-dimensional nonnegative vector $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ into Eq. (12) and taking the infinite limit when $K \to \infty$, an infinite factor model based on the GaP is given by

$$p(\boldsymbol{x}_n|\boldsymbol{\Theta}) = p\left(\boldsymbol{x}_n \middle| \sum_{k=1}^{\infty} \pi_k \omega_{nk} \boldsymbol{\phi}_k\right), \tag{13}$$

where π_k is a *global* weight of factor k and $\boldsymbol{\pi}$ does not need to sum to unity. To make $\boldsymbol{\pi}$ sparse, *i.e.*, to make fewer factors effective, a GaP prior GaP(G_0, α) with a base measure G_0 over a space Ω ($G_0(\Omega) = \gamma$ is the total mass of G_0 over Ω) and a concentration parameter α can be used as follows:

$$G \sim \text{GaP}(G_0, \alpha), \tag{14}$$

where G consists of infinitely many atom as follows:

$$\boldsymbol{\omega}_k, \boldsymbol{\phi}_k \sim \frac{1}{\gamma} G_0, \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\omega}_k, \boldsymbol{\phi}_k}, \tag{15}$$

While a SBP for generating $\boldsymbol{\pi}$ was recently derived [20], a CRP representation has not been proposed yet. Although a VB method based on the SBP representation was proposed for estimating the effective number of factors K^+ in matrix factorization [20], it has scarcely been used in practice due to its complexity. Another simpler representation of the GaP is weak-limit approximation that independently puts an extremely sparse gamma prior on each π_k as follows:

$$\pi_k \sim \text{Gamma}\left(\frac{\alpha\gamma}{K}, \alpha\right). \tag{16}$$

Since the GaP is obtained as the limit of Eq. (16) when $K \to \infty$, K is set to a sufficiently large number in practice in exchange for unnecessarily increase of computational cost.

A VB method based on Eq. (16) was proposed for GaP-NMF and applied to music signal analysis [9]. It is, however, often difficult to determine a threshold for gradually removing ineffective factors to estimate K^+. In addition, the results obtained by the weak-limit approximation tend to be sensitive to the truncation level.

2.2.2. Beta process

Introducing a binary matrix $\boldsymbol{Z} = \{z_{nk}\}_{n=1,k=1}^{N,K}$ into Eq. (12) and taking the infinite limit when $K \to \infty$, an infinite factor model based on the BP is given by

$$p(\boldsymbol{x}_n|\boldsymbol{\Theta}) = p\left(\boldsymbol{x}_n \middle| \sum_{k=1}^{\infty} z_{nk} \omega_{nk} \boldsymbol{\phi}_k\right), \tag{17}$$

where z_{nk} is a latent binary variable that indicates the presence or absence of factor k in sample n. To make $\boldsymbol{Z}$ sparse, a BP prior BP(G_0, α) with a base measure G_0 over a space Ω ($G_0(\Omega) = \gamma$) and a concentration parameter α is used with a Bernoulli process BeP(H) with a base measure H as follows:

$$H \sim \text{BP}(G_0, \alpha), \quad G_n \sim \text{BeP}(H), \tag{18}$$

where G and H over Ω can be explicitly written as

$$\boldsymbol{\omega}_k, \boldsymbol{\phi}_k \sim \frac{1}{\gamma} G_0, \tag{19}$$

$$H = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\omega}_k, \boldsymbol{\phi}_k}, \quad G_n = \sum_{k=1}^{\infty} z_{nk} \delta_{\boldsymbol{\omega}_k, \boldsymbol{\phi}_k}, \tag{20}$$

where $\pi_k \in [0, 1]$ indicates the probability of activating factor k and $z_{nk} \sim \text{Ber}(\pi_k)$. Since the infinite-dimensional vector $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{\infty}$ is extremely sparse, only a limited number of different factors are activated in $\boldsymbol{Z}$.

Several kinds of SBPs have been proposed for explicitly representing $\boldsymbol{\pi}$. Teh *et al.* [21], for example, proposed a SBP similar to Eq. (5) and derived a GS method that can estimate $\boldsymbol{\Theta}$ without truncated approximation by using slice sampling. Gupta *et al.* [22] proposed BP-NMF based on the same SBP representation. Paisley *et al.* [23, 24] proposed another SBP as the infinite limit of a finite model and derived a VB method. Another simpler representation of the BP is weak-limit approximation that independently puts an extremely sparse beta prior on each π_k as follows:

$$\pi_k \sim \text{Beta}\left(\frac{\alpha\gamma}{K}, \alpha\left(1 - \frac{\gamma}{K}\right)\right), \tag{21}$$

where K is set to a sufficiently large number as in the GaP. Liang *et al.* [7,8] proposed BP-NMF based on this representation and derived VB and GS methods with truncated approximation. A CRP-like representation of the BP is known as an Indian buffet process (IBP) [25].

3. DIRICHLET PROCESS PLCA

This section describes a nonparametric Bayesian extension of PLCA based on the Dirichlet process (DP) for audio source separation and then derives two kinds of learning methods based on VB and collapsed GS.

3.1. Data preparation

To use PLCA for audio source separation, the magnitude spectrogram of a mixture signal is regarded as a two-dimensional histogram of "sound quanta" in the time-frequency plane with N frames and M frequecy bins, as explained in Section 2.1.4. We redefine $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^I$ as a set of observed variables and let $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^I$ be a set of the corresponding latent variables, where I is the total number of sound quanta, *i.e.*, the sum of the values of magnitude over the time-frequency plane. Since each quantum i is located at a time-frequency bin, $\boldsymbol{x}_i$ is represented as an NM-dimensional one-hot vector that takes 1 in a dimension corresponding to the time-frequency location and 0 in the other dimensions. In addition, each quantum i is assumed to be generated from one of K sources ($K \to \infty$), $\boldsymbol{z}_i$ is represented as a K-dimensional one-hot vector.

To make the observed data $\boldsymbol{X}$, we need to quantize the value of magnitude at each time-frequency bin according to an appropriate resolution. In practice, PLCA is found to work by scaling and rounding the magnitude spectrogram such that the average of magnitude is around 1. If the average of magnitude is set to a larger value, the number of observations I takes a larger value, which affects the posterior uncertanty of the parameters of PLCA in Bayesian estimation. This is more problematic for nonparametric Bayesian PLCA because the effective number of sources, K^+, is considered to increase logarithmically according to the increase of observed data. To solve this problem, we could use a method of optimizing the resolution of magnitude quantization that was originally proposed for KL-NMF based on the discrete Poisson distribution [26].

3.2. Model formulation

The probabilistic model of PLCA for two-dimensional data is given by Eq. (10), where $p(k)$ represents the mixing ratios of K sources, and $p(n|k)$ and $p(m|k)$ represent the time and frequency distributions of source k, respectively. The parameters of these categorical distributions are given by $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$, $\boldsymbol{\phi}_k = \{\phi_{kn}\}_{n=1}^N$, and $\boldsymbol{\theta}_k = \{\theta_{km}\}_{m=1}^M$. The probabilistic generative model of $\boldsymbol{Z}$ and $\boldsymbol{X}$ (the complete likelihood function of $\boldsymbol{\pi}$, $\boldsymbol{\phi}$, and $\boldsymbol{\theta}$ for $\boldsymbol{Z}$ and $\boldsymbol{X}$) is thus defined as follows:

$$p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{i=1}^{I}\prod_{k=1}^{K}\pi_k^{z_{ik}}, \tag{22}$$

$$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\phi},\boldsymbol{\theta}) = \prod_{i=1}^{I}\prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{k=1}^{K}(\phi_{kn}\theta_{km})^{x_{inm}z_{ik}}. \tag{23}$$

To let K go to infinity, we use a DP prior. There are two major constructions of the DP (see Sections 2.1.1 and 2.1.2). Using a SBP and conjugate Dirichlet priors, a nonparametric Baysian model can be formulated as follows:

$$\boldsymbol{\pi} \sim \mathrm{SBP}(\alpha), \quad \boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta), \quad \boldsymbol{\theta}_k \sim \mathrm{Dir}(\gamma), \tag{24}$$

where α, β, and γ are hyperparameters. The SBP representation is convenient for deriving the VB method.

Using a CRP, *i.e.*, marginalizing $\boldsymbol{\pi}$ out analytically, we obtain another representation as follows:

$$\boldsymbol{Z} \sim \mathrm{CRP}(\alpha), \quad \boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta), \quad \boldsymbol{\theta}_k \sim \mathrm{Dir}(\gamma). \tag{25}$$

Note that $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can also be marginalized out for efficient Bayesian inference (only $\boldsymbol{Z}$ is considered). The CRP representation is convenient when the (collapsed) GS method is used.

3.3. Variational Bayes

Using a Bayesian model specified by Eqs. (22), (23), and (24), we aim to calculate the posterior distribution $p(\boldsymbol{Z},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{X})$, where $\boldsymbol{\eta}$ is considered instead of $\boldsymbol{\pi}$ based on Eq. (5). Since the true posterior is analytically intractable, it is approximated as a factorizable variational posterior distribution $q(\boldsymbol{Z},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\theta}) = q(\boldsymbol{Z})q(\boldsymbol{\eta})q(\boldsymbol{\phi})q(\boldsymbol{\theta})$ such that the KL divergence from $q(\boldsymbol{Z},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\theta})$ to $p(\boldsymbol{Z},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{X})$ is minimized, *i.e.*, the lower bound of the log-evidence $p(\boldsymbol{X})$ is maximized. In each iteration, $q(\boldsymbol{Z})$ is updated as follows:

$$q(\boldsymbol{Z}) \propto \exp(\mathbb{E}_{q(\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\theta})}[\log p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\theta})]). \tag{26}$$

Similarly, $q(\boldsymbol{\eta})$, $q(\boldsymbol{\phi})$, and $q(\boldsymbol{\theta})$ can also be updated alternately. Each posterior is found to be the same type of the prior distribution because of the conjugacy.

First, $q(\boldsymbol{Z})$ is given by a categorical distribution as follows:

$$q(\boldsymbol{z}_i) = \mathrm{Categorical}(\boldsymbol{\zeta}_i), \tag{27}$$

where $\zeta_{ik} = \frac{\rho_{ik}}{\sum_{k'=1}^K \rho_{ik'}}$ and ρ_{ik} is given by

$$\log\rho_{ik} = \mathbb{E}[\log\eta_k] + \sum_{k'=1}^{k-1}\mathbb{E}[\log(1-\eta_{k'})] + \sum_{nm} x_{inm}\mathbb{E}[\log\phi_{kn}] + \sum_{nm} x_{inm}\mathbb{E}[\log\theta_{km}]. \tag{28}$$

Then, $q(\boldsymbol{\eta})$ is given by

$$q(\eta_k) = \mathrm{Beta}\left(1 + \sum_i \mathbb{E}[z_{ik}], \alpha + \sum_i\sum_{k'=k+1}^{K}\mathbb{E}[z_{ik'}]\right), \tag{29}$$

Finally, $q(\boldsymbol{\phi})$ and $q(\boldsymbol{\theta})$ are given by

$$q(\boldsymbol{\phi}_k) = \mathrm{Dir}(\boldsymbol{\lambda}_k), \quad q(\boldsymbol{\theta}_k) = \mathrm{Dir}(\boldsymbol{\omega}_k), \tag{30}$$

where $\boldsymbol{\lambda}_k$ and $\boldsymbol{\omega}_k$ are given by

$$\lambda_{kn} = \beta + \sum_{im} x_{inm}\mathbb{E}[z_{ik}], \quad \omega_{km} = \gamma + \sum_{in} x_{inm}\mathbb{E}[z_{ik}]. \tag{31}$$

As in GaP-NMF, K is initialized as a sufficiently large number and uneffective sources are gradually removed in each iteration.

3.4. Collapsed Gibbs sampling

Formulating a Bayesian model given by Eqs. (22), (23), and (25) and marginalizing out $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by leveraging the conjugacy between the Dirichlet and categorical distributions, we aim to draw samples from the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X})$ because at most N different classes appear in $\boldsymbol{Z}$ (the number of different classes K^+ is usually much fewer than N). This enables us to avoid dealing with infinitely many parameters on computers. We use collapsed Gibbs sampling for updating each $\boldsymbol{z}_i$ in a random order according to the following conditional distribution:

$$p(\boldsymbol{z}_i|\boldsymbol{Z}_{\neg i},\boldsymbol{X}) \propto p(\boldsymbol{z}_i,\boldsymbol{x}_i|\boldsymbol{Z}_{\neg i},\boldsymbol{X}_{\neg i}), \tag{32}$$

where $X\neg x$ indicates a set of parameters X except for x. When $\boldsymbol{z}_i$ is updated, K^+ might be decremented, incremented, or unchanged. More specifically, the probability that $\boldsymbol{z}_i$ is generated from "existing" source k ($z_{ik}=1$) and $\boldsymbol{x}_i$ is then located at time n and frequency bin m ($x_{inm}=1$) is given by

$$p(z_{ik}=1, x_{inm}=1|\boldsymbol{Z}_{\neg i},\boldsymbol{X}_{\neg i}) \propto \frac{\sum_{i'\neq i} z_{i'k}}{I-1+\alpha} \frac{\sum_m\sum_{i'\neq i} x_{i'nm}z_{i'k}+\beta}{\sum_{i'\neq i} z_{i'k}+\beta N} \frac{\sum_n\sum_{i'\neq i} x_{i'nm}z_{i'k}+\gamma}{\sum_{i'\neq i} z_{i'k}+\gamma M}. \tag{33}$$

On the other hand, the probability that $\boldsymbol{z}_i$ is generated from a new source k_{new} ($z_{ik_{\text{new}}}=1$) and $\boldsymbol{x}_i$ is located at time n and frequency bin m ($x_{inm}=1$) is given by

$$p(z_{ik_{\text{new}}}=1, x_{inm}=1|\boldsymbol{Z}_{\neg i},\boldsymbol{X}_{\neg i}) \propto \frac{\alpha}{I-1+\alpha}\frac{1}{MN}. \tag{34}$$

4. EVALUATION

This section reports a comparative experiment evaluatining the capability of DP-PLCA in estimation of the number of sources K^+.

4.1. Experimental conditions

We used three mixture signals each of which was synthesized using the sounds of piano (011PFNOM), electric guitar (131EGLPM), or clarinet (311CLNOM) recorded in the RWC Music Database: Musical Instrument Sound [27]. Each signal (14 s) was made by concatenating seven 2-s isolated or mixture sounds (C4, E4, G4, C4+E4, C4+G4, E4+G4, and C4+E4+G4). The sampling rate was 16 kHz. Each mixture signal was expected to be separated into three sources corresponding to C4, E4, and G4 ($K^+=3$). The short-time Fourier transform (STFT) with a Gaussian window was performed with a window length of 512 pts and a shifting interval of 160 pts. The magnitude spectrogram of each signal was represented as a nonnegative matrix with $N=1400$ and $M=257$ and scaled such that the average value of magnitude over the time-frequency plane, μ, was equal to 1 or 10. We tested the VB and GS methods for learning the DP-PLCA model. For comparison, we tested GaP-NMF based on the Poisson distribution (GaP extension of KL-NMF, Section 2.2.1)

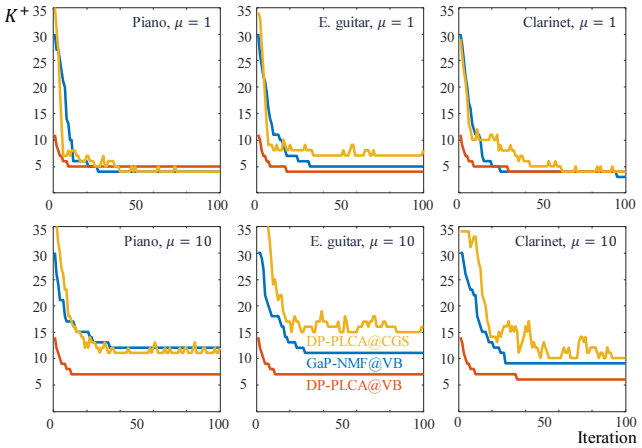

Fig. 5. Experimental results of K^+ estimation.

based on the VB method. The truncation level of the VB methods was set to 30 and the GS method started from 30 sources. The performance of each method was evaluated in terms of accuracy of K^+ estimation. When $K^+ = 3$, these methods achieved comaprable source separation performance.

4.2. Experimental results

Fig. 5 shows the experimental results. When $\mu = 1$ (*i.e.,* the total number of sound quanta was $I = NM$), all methods converged to a reasonable solution around $K^+ = 4$, where an extra source represents a noise spectrum (atack sound). When $\mu = 10$, all methods tended to overestimate K^+. Although this is a natural behavior of nonparametric Bayesian models, the essential complexity of $\boldsymbol{X}$ remains the same. DP-PLCA based on the VB method was found to be less sensitive to the apparent data size. The computational cost of the VB method for DP-PLCA was much smaller than that for GaP-NMF. Note that the computational costs of the VB methods are independent from the data size I, that of the GS method is linearly increased. One solution is to use non-collapsed GS that can adaptively truncate $\boldsymbol{\pi}$ without any approximation by using slice sampling. This enables parall computation as in the VB methods.

5. CONCLUSION

This paper presented a nonparametric Bayesian extension of PLCA called DP-PLCA to estimate the number of sources in audio source separation. One of the major contributions of this paper is to clarify the essential difference between two major matrix factorization techniques, NMF (factor model) and PLCA (mixture model), which are often mistakenly considered to be *always* identical, in terms of probabilistic modeling. This is a reason why the DP is used with PLCA while the GaP or BP is used with NMF. We derived the learning methods based on VB and GS and confirmed that these methods outperformed GaP-NMF in terms of the capability of autonomous model complexity control.

6. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2004.

[2] P. Smaragdis et al., "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Sig. Proc. Mag.*, vol. 31, no. 3, pp. 66–75, 2014.

[3] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence and Neuroscience*, vol. 2008, pp. 1–8, 2008.

[4] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," *WASPAA*, 2003.

[5] C. Févotte et al., "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[6] T. Hofmann, "Learning the similarity of documents: An information-geometric approach to document retrieval and categorization," *NIPS*, 2000, pp. 914–920.

[7] D. Liang, M. Hoffman, and D. Ellis, "Beta process sparse nonnegative matrix factorization for music," *ISMIR*, 2013.

[8] D. Liang and M. D. Hoffman, "Beta process non-negative matrix factorization with stochastic structured mean-field variational inference," *NIPS Workshop*, 2010.

[9] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," *ICML*, 2010.

[10] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[11] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[12] T. Ferguson, "Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[13] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[15] Y. W. Teh et al., "Hierarchical Dirichlet processes," *JASA*, vol. 101, pp. 1566–1581, 2006.

[16] B. Fuentes, R. Badeau, and G. Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription," *IEEE TASLP*, vol. 21, no. 9, pp. 1854–1866, 2013.

[17] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model," *JASA*, vol. 133, no. 3, pp. 1727–1741, 2013.

[18] M. Zhou et al., "Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Trans. on Image Proc.*, vol. 21, no. 1, pp. 130–144, 2012.

[19] H. P. Dang and P. Chainais, "A Bayesian non parametric approach to learn dictionaries with adapted numbers of atoms," *MLSP*, 2015, pp. 1–6.

[20] A. Roychowdhury and B. Kulis, "Gamma processes, stick-breaking, and variational inference," *AISTATS*, 2015, pp. 800–808.

[21] Y. W. Teh et al., "Stick-breaking construction for the Indian buffet process," *AISTATS*, 2007, pp. 556–563.

[22] S. K. Gupta et al., "A nonparametric Bayesian Poisson gamma model for count data," *ICPR*, 2012, pp. 1815–1818.

[23] J. Paisley et al., "A stick-breaking construction of the beta process," *ICML*, 2010, pp. 847–854.

[24] J. Paisley, L. Carin, and D. Blei, "Variational inference for stick-breaking beta process priors," *ICML*, 2011, pp. 889–896.

[25] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," *NIPS*, 2006, pp. 475–482.

[26] M. D. Hoffman, "Poisson-uniform nonnegative matrix factorization," *ICASSP*, 2012, pp. 5361–5364.

[27] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," *ISMIR*, 2002, pp. 287–288.