Infinite Positive Semidefinite Tensor Factorization for Source Separation of Mixture Signals

Kazuyoshi Yoshii K.YOSHII@AIST.GO.JP
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568, Japan

Ryota Tomioka TOMIOKA@MIST.I.U-TOKYO.AC.JP
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Daichi Mochihashi DAICHI@ISM.AC.JP
The Institute of Statistical Mathematics (ISM), 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

Masataka Goto M.GOTO@AIST.GO.JP
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568, Japan

Abstract

This paper presents a new class of tensor factorization called *positive semidefinite tensor factorization* (PSDTF) that decomposes a set of positive semidefinite (PSD) matrices into the convex combinations of fewer PSD basis matrices. PSDTF can be viewed as a natural extension of *nonnegative matrix factorization*. One of the main problems of PSDTF is that an appropriate number of bases should be given in advance. To solve this problem, we propose a nonparametric Bayesian model based on a gamma process that can instantiate only a limited number of necessary bases from the infinitely many bases assumed to exist. We derive a variational Bayesian algorithm for closed-form posterior inference and a multiplicative update rule for maximum-likelihood estimation. We evaluated PSDTF on both synthetic data and real music recordings to show its superiority.

1. Introduction

Matrix factorization (MF) has recently been an active research topic in the field of machine learning. Given a matrix $\boldsymbol{X} \in \mathbb{R}^{M \times N}$ as observed data, the objective is to find a low-rank approximation $\boldsymbol{X} \approx \boldsymbol{AB}$ where $\boldsymbol{A} \in \mathbb{R}^{M \times K}$, $\boldsymbol{B} \in \mathbb{R}^{K \times N}$, and $K \ll \min\{M, N\}$. This problem often arises in many application fields. Candidates of $\boldsymbol{X}$ include a user-item rating matrix in collaborative filtering (Salakhutdinov & Mnih, 2008), a set of face images in image processing (Lee & Seung, 2000), and a time-frequency spectrogram in audio processing (Smaragdis & Brown, 2003). Many variants of MF have been proposed by using various measures on the reconstruction error $\mathcal{D}(\boldsymbol{X}|\boldsymbol{AB})$ and imposing constraints on $\boldsymbol{A}$ and $\boldsymbol{B}$. In terms of probabilistic modeling, a specific model is defined by a likelihood function $p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{B})$ and prior distributions $p(\boldsymbol{A})$ and $p(\boldsymbol{B})$.

One of the popular classes of MF is nonnegative matrix factorization (NMF), in which all elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ must be no less than zero. This constraint reflects the fact that some physical quantities, *e.g.*, pixel brightness and signal energy, cannot be negative. A typical way to impose this constraint is to place *element-wise* gamma priors on $\boldsymbol{A}$ and $\boldsymbol{B}$ (Cemgil, 2009). $\mathcal{D}(\boldsymbol{X}|\boldsymbol{AB})$ has often been defined in an *element-wise* manner by using the Bregman divergence (Bregman, 1967), which includes as special cases the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) and the Itakura-Saito (IS) divergence (Itakura & Saito, 1968). The assumption underlying the likelihood $p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{B})$ is that each element of $\boldsymbol{X}$ is *independently* Poisson distributed in KL-NMF or *independently* exponentially distributed in IS-NMF.

In audio analysis, IS-NMF is more suitable for decomposing a power spectrogram $\boldsymbol{X}$ over M frequency bins and N frames as the product of sound-source power spectra (K columns of $\boldsymbol{A}$) and the corresponding temporal activations (K rows of $\boldsymbol{B}$) (Févotte et al., 2009). IS-NMF is theoretically justified if the frequency bins of source spectra are independent. Note that the au-

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

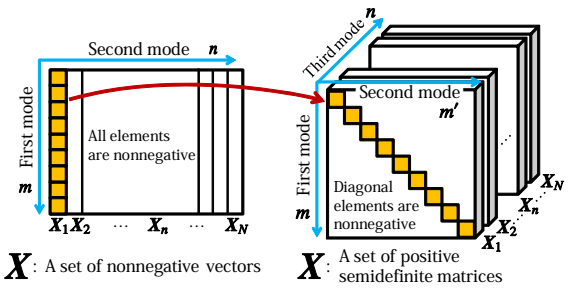

Figure 1. Comparison between IS-NMF and LD-PSDTF.

dio signals of pitched instruments have clear periodicities, *i.e.*, those signals are highly autocorrelated at certain time lags. However, the short-term Fourier transform (STFT) is unable to perfectly decorrelate the frequency components forming harmonic structures. A similar problem arises in electroencephalogram (EEG) analysis (Lee et al., 2006), in which cross-correlations between multichannel signals recorded at different positions of the head are usually ignored. This indicates that it is not appropriate to place gamma priors on $\boldsymbol{A}$ and define $\mathcal{D}(\boldsymbol{X}|\boldsymbol{AB})$ in an *element-wise* manner.

To solve this problem, we propose a new class of tensor factorization called positive semidefinite tensor factorization (PSDTF). As NMF decomposes N nonnegative vectors (a matrix) as the conic sums of K nonnegative vectors, PSDTF decomposes N PSD matrices (a tensor) as the conic sums of K PSD matrices. As shown in Figure 1, each nonnegative vector is embeded into a PSD matrix that represents the covariance structure of the multivariate elements. We thus place *matrix-wise* Wishart priors on the basis matrices. In this paper the reconstruction error is defined by using a kind of the Bregman matrix divergence called the log-determinant (LD) divergence (Kulis et al., 2009), also in a *matrix-wise* manner. This implies that each slice of the observed tensor is assumed to have a Wishart likelihood. Since the resulting LD-PSDTF is a natural extension of IS-NMF, an inherited problem is that the number of bases, K, should be given in advance.

To estimate an appropriate number of basis matrices, we propose a nonparametric Bayesian model of LD-PSDTF similar to one of IS-NMF (Hoffman et al., 2010). Although the Wishart prior-Wishart likelihood hierarchy does not satisfy the conjugacy condition, we can derive an elegant variational algorithm for closed-form posterior inference. A multiplicative update rule can also be used for maximum-likelihood estimation. In addition, we reveal that IS-NMF in the frequency domain approximates LD-PSDTF in the time domain that can consider periodic covariance structures of the audio signals. This explains why IS-NMF works well for music power-spectrogram decomposition.

2. Gamma Process Positive Semidefinite Tensor Factorization

We propose a probabilistic model of positive semidefinite tensor factorization (PSDTF) and derive its nonparametric Bayesian extension that in theory allows observed data (*e.g.*, a music signal) to contain an infinite number of latent bases (*e.g.*, sound sources) by using the gamma process. An *effective* number of bases required for representing the observed data can be efficiently estimated in a data-driven manner. We then discuss how PSDTF is related to matrix factorization and how it is applied to music signal analysis.

2.1. Problem Specification

We will formalize the problem. Suppose we have as observed data a three-mode tensor $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N] \in \mathbb{R}^{M\times M\times N}$, where each slice $\boldsymbol{X}_n \in \mathbb{R}^{M\times M}$ is a real symmetric positive semidefinite (PSD) matrix. Although PSDTF can be defined even if $\boldsymbol{X}_n \in \mathbb{C}^{M\times M}$ is a complex Hermitian PSD matrix, we focus on the case of $\boldsymbol{X}_n \in \mathbb{R}^{M\times M}$ for explanatory simplicity.

The goal of factorization is to approximate each PSD matrix $\boldsymbol{X}_n$ by a convex combination of PSD matrices $\{\boldsymbol{V}_k\}_{k=1}^K$ (K bases) as follows:

$$\boldsymbol{X}_n \approx \sum_{k=1}^{K} \theta_k h_{kn} \boldsymbol{V}_k \equiv \boldsymbol{Y}_n, \tag{1}$$

where $\theta_k \geq 0$ is a *global* weight shared over all N slices and $h_{kn} \geq 0$ is a *local* weight specific to the n-th slice. Eq. (1) can also be represented as $\boldsymbol{X} \approx \sum_{k=1}^K \theta_k \boldsymbol{h}_k \otimes \boldsymbol{V}_k \equiv \boldsymbol{Y}$, where $\otimes$ indicates the Kronecker product.

To evaluate the reconstruction error between PSD matrices $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$, we propose to use a Bregman matrix divergence (Bregman, 1967) defined as follows:

$$\begin{aligned}&\mathcal{D}_\phi(\boldsymbol{X}_n|\boldsymbol{Y}_n)\\&= \phi(\boldsymbol{X}_n) - \phi(\boldsymbol{Y}_n) - \mathrm{tr}\big(\nabla\phi(\boldsymbol{Y}_n)^T(\boldsymbol{X}_n - \boldsymbol{Y}_n)\big),\end{aligned} \tag{2}$$

where ϕ is a strictly convex matrix function. In this paper we focus on the log-determinant (LD) divergence ($\phi(\boldsymbol{Z}) = -\log|\boldsymbol{Z}|$) (Kulis et al., 2009) given by

$$\begin{aligned}&\mathcal{D}_{\mathrm{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n)\\&= -\log\left|\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}\right| + \mathrm{tr}\left(\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}\right) - M.\end{aligned} \tag{3}$$

This divergence is always nonnegative and is zero if and only if $\boldsymbol{X}_n = \boldsymbol{Y}_n$ holds. A well-known special case when $M = 1$ is the Itakura-Saito (IS) divergence over nonnegative numbers (Itakura & Saito, 1968) given by $\mathcal{D}_{\mathrm{IS}}(x|y) = -\log(x/y) + x/y - 1$ and is often used in signal processing. Sivalingam et al. (2010) formalized a similar tensor factorization problem that uses $\mathcal{D}_{\mathrm{LD}}(\boldsymbol{Y}_n|\boldsymbol{X}_n)$ as a cost function.

Our goal here is to estimate unknown variables $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_K]^T \in \mathbb{R}^K$, $\boldsymbol{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_K] \in \mathbb{R}^{N \times K}$, and $\boldsymbol{V} = [\boldsymbol{V}_1, \cdots, \boldsymbol{V}_K] \in \mathbb{R}^{M \times M \times K}$ such that the cost function $\mathcal{C}(\boldsymbol{X}|\boldsymbol{Y}) = \sum_n \mathcal{D}_{\mathrm{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ is minimized. Note that our model imposes the nonnegativity constraint on $\boldsymbol{\theta}$ and $\boldsymbol{H}$ and the positive semidefiniteness constraint on $\boldsymbol{V}$. We call this model LD-PSDTF.

2.2. Probabilistic Formulation

We explain Bayesian treatment of LD-PSDTF defined by Eq. (1) in terms of probabilistic modeling.

2.2.1. Formulating a Probabilistic Model

We first formulate a finite model based on a fixed number of bases by specifying prior distributions on $\boldsymbol{\theta}$, $\boldsymbol{H}$, and $\boldsymbol{V}$ and a likelihood function of $\boldsymbol{X}$. In this model we assume $\theta_k = 1$ because the effect of θ_k can be compensated by adjusting the scale of $\boldsymbol{h}_k$.

Since h_{kn} is nonnegative ($h_{kn} \geq 0$) and $\boldsymbol{V}_k$ is PSD ($\boldsymbol{V}_k \geq \boldsymbol{0}$), a natural choice is to place gamma and Wishart priors on h_{kn} and $\boldsymbol{V}_k$ as follows:

$$h_{kn} \sim \mathcal{G}(a_0, b_0), \tag{4}$$

$$\boldsymbol{V}_k \sim \mathcal{W}(\nu_0, \boldsymbol{V}_0), \tag{5}$$

where a_0 and b_0 are the shape and rate parameters of the gamma distribution and ν_0 and $\boldsymbol{V}_0$ are the degree of freedom (DOF) and scale matrix of the Wishart distribution.

We then assume PSD matrices $\{\nu \boldsymbol{X}_n\}_{n=1}^N$ to be independently Wishart distributed as follows:

$$\nu \boldsymbol{X}_n | \boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V} \sim \mathcal{W}\left(\nu, \sum_{k=1}^K \theta_k h_{kn} \boldsymbol{V}_k\right), \tag{6}$$

where ν is a DOF of the Wishart distribution. Note that $\mathbb{E}[\boldsymbol{X}_n] = \boldsymbol{Y}_n$ and $\mathbb{M}[\boldsymbol{X}_n] = \frac{\nu - M - 1}{\nu} \boldsymbol{Y}_n$, where $\mathbb{M}$ means the mode. When $\nu \gg M$, $\mathbb{M}[\boldsymbol{X}_n] \approx \boldsymbol{Y}_n$ holds. When $\nu < M$, $\boldsymbol{X}_n$ is rank deficient. If $M = \nu = 1$, the distribution reduces to an exponential distribution. The log-likelihood of $\boldsymbol{X}_n$ is given by

$$\log p(\boldsymbol{X}_n|\boldsymbol{Y}_n) = C(\nu) + \frac{\nu - M - 1}{2} \log |\boldsymbol{X}_n| - \frac{\nu}{2} \log |\boldsymbol{Y}_n| - \frac{\nu}{2} \mathrm{tr}\left(\boldsymbol{X}_n \boldsymbol{Y}_n^{-1}\right), \tag{7}$$

where $C(\nu)$ is a constant term depending only on ν and the second term can also be considered to be constant because $\boldsymbol{X}_n$ is the observed data. Therefore, the maximization of the likelihood $p(\boldsymbol{X}|\boldsymbol{Y}) = \prod_n p(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ with respect to $\boldsymbol{Y}$ is equivalent to the minimization of the cost function $\mathcal{C}(\boldsymbol{X}|\boldsymbol{Y}) = \sum_n \mathcal{D}_{\mathrm{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ (compare Eq. (7) with Eq. (3)).

Consequently, a Bayesian model of LD-PSDTF is defined by Eqs. (4), (5), and (6). Given the data $\boldsymbol{X}$, our goal is to calculate a posterior distribution $p(\boldsymbol{H}, \boldsymbol{V}|\boldsymbol{X})$ over unknown variables $\boldsymbol{H}$ and $\boldsymbol{V}$.

2.2.2. Taking the Infinite Limit

To overcome the limitation that the number of bases K should be specified in advance, we leverage Bayesian nonparametrics for taking the infinite limit of Eq. (6) as K diverges to infinity. Given the data $\boldsymbol{X}$, an *effective* number of bases should be estimated in a data-driven manner. We thus aim to learn a sparse infinite-dimensional weight vector $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_\infty]^T$ as proposed by Hoffman et al. (2010).

We place a gamma process (GaP) prior on $\boldsymbol{\theta}$ in a so-called weak-approximation manner as follows:

$$\theta_k \sim \mathcal{G}(\alpha c/K, \alpha), \tag{8}$$

where α and c are positive numbers, $\mathbb{E}_{\mathrm{prior}}[\theta_k] = c/K$, and $\mathbb{E}_{\mathrm{prior}}[\sum_k \theta_k] = c$. When the truncation level K goes to infinity, the vector $\boldsymbol{\theta}$ approximates an infinite-dimensional discrete measure $\boldsymbol{G}$ that is stochastically drawn from the GaP over a space $\boldsymbol{\Theta}$ as follows:

$$\boldsymbol{G} \sim \mathrm{GaP}(\alpha, \boldsymbol{G}_0), \tag{9}$$

where α is called a concentration parameter and $\boldsymbol{G}_0$ a base measure. In our model we assumed that $\boldsymbol{G}_0$ is a uniform measure such that $\boldsymbol{G}_0(\boldsymbol{\Theta}) = c$. The *effective* number of elements, K^+, such that $\theta_k > \epsilon$ for some number $\epsilon > 0$ is almost surely finite. If we set K to be sufficiently larger than α, only a few of the K elements of $\boldsymbol{\theta}$ will be substantially greater than zero.

A nonparametric Bayesian model of GaP-LD-PSDTF is defined by Eqs. (4), (5), (6), and (8) with a large truncation level K. Given the data $\boldsymbol{X}$, our goal is to calculate a posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V}|\boldsymbol{X})$ and estimate the value of K^+ at the same time.

2.3. "Augmented" Matrix Factorization

We show here that LD-PSDTF naturally emerges from the standard problem of matrix factorization. Suppose we have a set of N samples $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_N] \in \mathbb{R}^{M \times N}$ as observed data, where $\hat{\boldsymbol{x}}_n \in \mathbb{R}^M$ is a feature vector of the n-th sample. Although the case of $\hat{\boldsymbol{x}}_n \in \mathbb{C}^M$ can be dealt with, as in Section 2.1 we here discuss the case of $\hat{\boldsymbol{x}}_n \in \mathbb{R}^M$. In signal processing, for example, a local signal $\boldsymbol{s}_n \in \mathbb{R}^M$ in the n-th short window (called a frame) is often regarded as $\hat{\boldsymbol{x}}_n$ (Figure 2). Alternatively, $\hat{\boldsymbol{x}}_n$ can be a complex spectrum $\boldsymbol{c}_n = \boldsymbol{F}\boldsymbol{s}_n \in \mathbb{C}^M$ at the n-th frame, where $\boldsymbol{F} \in \mathbb{C}^{M \times M}$ is the unitary discrete Fourier transform matrix.

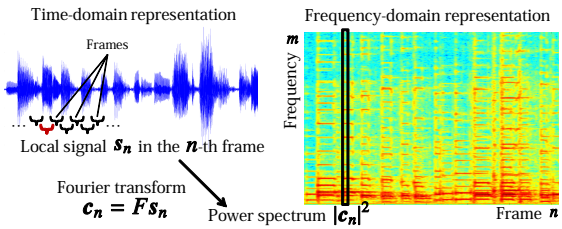

Figure 2. Different representations of a music audio signal.

The goal is to discover a limited number of bases having characteristic structures (*e.g.*, instrument sounds) from the observed data $\hat{\boldsymbol{X}}$ (*e.g.*, music audio signal), *i.e.*, to decompose each sample $\hat{\boldsymbol{x}}_n$ into a linear sum of K *variable* bases $\{\hat{\boldsymbol{w}}_{kn}\}_{k=1}^K$ as follows:

$$\hat{\boldsymbol{x}}_n = \sum_{k=1}^K \hat{\boldsymbol{x}}_{nk} = \sum_{k=1}^K \hat{\theta}_k \hat{h}_{kn} \hat{\boldsymbol{w}}_{kn}, \tag{10}$$

where $\hat{\theta}_k$ is a *global* coefficient of the k-th basis, $\hat{h}_{kn}$ is a *local* coefficient of the k-th basis, and $\hat{\boldsymbol{x}}_{nk} = \hat{\theta}_k \hat{h}_{kn} \hat{\boldsymbol{w}}_{kn}$ is the k-th component in $\hat{\boldsymbol{x}}_n$. Those variables are allowed to take any real values. If we assume that $\hat{\boldsymbol{w}}_{kn}$ is a *fixed* basis such that $\hat{\boldsymbol{w}}_{kn}$ is equal to $\hat{\boldsymbol{w}}_k$ for any n and define some symbols as $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \cdots, \hat{\theta}_K]^T \in \mathbb{R}^K$, $\hat{\boldsymbol{W}} = [\hat{\boldsymbol{w}}_1, \cdots, \hat{\boldsymbol{w}}_k] \in \mathbb{R}^{M \times K}$, and $\hat{\boldsymbol{H}} = [\hat{\boldsymbol{h}}_1, \cdots, \hat{\boldsymbol{h}}_K] \in \mathbb{R}^{N \times K}$, Eq. (10) can be simply written as follows:

$$\hat{\boldsymbol{X}} = \hat{\boldsymbol{W}} \mathrm{diag}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{H}}^T, \tag{11}$$

where $\mathrm{diag}(\boldsymbol{z})$ means a diagonal matrix having a vector $\boldsymbol{z}$ as its diagonal elements. If $\mathrm{diag}(\hat{\boldsymbol{\theta}})$ is an identity matrix ($\hat{\boldsymbol{\theta}} = \mathbf{1}$), Eq. (11) reduces to the standard problem of matrix factorization given by $\hat{\boldsymbol{X}} = \hat{\boldsymbol{W}} \hat{\boldsymbol{H}}^T$. The optimal values of $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{W}}$, and $\hat{\boldsymbol{H}}$ depend on what kinds of constraints are placed on those variables.

2.3.1. Formulating a Probabilistic Model

We aim to formulate a Bayesian model of Eq. (10). A key feature is to consider essential correlations between M elements of basis $\hat{\boldsymbol{w}}_{kn}$. A natural choice is to put a multivariate Gaussian prior on $\hat{\boldsymbol{w}}_{kn}$ as follows:

$$\hat{\boldsymbol{w}}_{kn} \sim \mathcal{N}(\mathbf{0}, \hat{\boldsymbol{V}}_k), \tag{12}$$

where $\hat{\boldsymbol{V}}_k \in \mathbb{R}^{M \times M}$ is a *full* covariance matrix. In general, the Gaussian mean is set to a zero vector. For example, an audio signal is recorded as real numbers distributed on both sides of zero (see Figure 2).

The linear relationship $\hat{\boldsymbol{x}}_{nk} = \hat{\theta}_k \hat{h}_{kn} \hat{\boldsymbol{w}}_{kn}$ and Eq. (12) lead to a likelihood of $\hat{\boldsymbol{x}}_{nk}$ as follows:

$$\hat{\boldsymbol{x}}_{nk} | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}} \sim \mathcal{N}(\mathbf{0}, \hat{\theta}_k^2 \hat{h}_{kn}^2 \hat{\boldsymbol{V}}_k). \tag{13}$$

Then, using the reproducing property of the Gaussian and the linear relationship $\hat{\boldsymbol{x}}_n = \sum_{k=1}^K \hat{\boldsymbol{x}}_{nk}$, we get a likelihood of $\hat{\boldsymbol{x}}_n$ as follows:

$$\hat{\boldsymbol{x}}_n | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}} \sim \mathcal{N}\left(\mathbf{0}, \sum_{k=1}^K \hat{\theta}_k^2 \hat{h}_{kn}^2 \hat{\boldsymbol{V}}_k\right). \tag{14}$$

If we assume that $\theta_k = \hat{\theta}_k^2 \geq 0$, $h_{kn} = \hat{h}_{kn}^2 \geq 0$, $\boldsymbol{V}_k = \hat{\boldsymbol{V}}_k \geq \mathbf{0}$, and $\boldsymbol{X}_n = \hat{\boldsymbol{x}}_n \hat{\boldsymbol{x}}_n^T \geq \mathbf{0}$, Eq. (14) recovers Eq. (7) when $\nu = 1$ except for constant terms. We can put the same priors as Eqs. (4), (5), and (8).

This is a special case of the general LD-PSDTF model in which each $\boldsymbol{X}_n$ is restricted to a *rank-1* PSD matrix ($\boldsymbol{X}_n = \hat{\boldsymbol{x}}_n \hat{\boldsymbol{x}}_n^T$). In general, the DOF of $\boldsymbol{X}$ can be larger than MN because each $\boldsymbol{X}_n$ is allowed to take *any* PSD matrix. In this section, the DOF of $\boldsymbol{X}$ is MN because $\boldsymbol{X}_n$ is just an augmented representation of $\hat{\boldsymbol{x}}_n$.

2.3.2. Estimating the Latent Components

In many applications such as source separation, latent component $\hat{\boldsymbol{x}}_{nk}$ is of main interest. One might consider it necessary to calculate $\hat{\boldsymbol{x}}_{nk} = \hat{\theta}_k \hat{h}_{kn} \hat{\boldsymbol{w}}_{kn}$. Instead, marginalizing out $\hat{\boldsymbol{w}}_{kn}$ gives the posterior of $\hat{\boldsymbol{x}}_{nk}$ as a Gaussian whose mean and covariance are

$$\mathbb{E}[\hat{\boldsymbol{x}}_{nk} | \hat{\boldsymbol{x}}_n, \boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V}] = \boldsymbol{Y}_{nk} \boldsymbol{Y}_n^{-1} \hat{\boldsymbol{x}}_n, \tag{15}$$

$$\mathbb{V}[\hat{\boldsymbol{x}}_{nk} | \hat{\boldsymbol{x}}_n, \boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V}] = \boldsymbol{Y}_{nk} - \boldsymbol{Y}_{nk} \boldsymbol{Y}_n^{-1} \boldsymbol{Y}_{nk}, \tag{16}$$

where $\boldsymbol{Y}_{nk} = \theta_k h_{kn} \boldsymbol{V}_k$ and $\boldsymbol{Y}_n = \sum_{k=1}^K \boldsymbol{Y}_{nk}$ are PSD matrices. For a Bayesian treatment, we need to calculate $\mathbb{E}[\hat{\boldsymbol{x}}_{nk} | \hat{\boldsymbol{x}}_n]$ and $\mathbb{V}[\hat{\boldsymbol{x}}_{nk} | \hat{\boldsymbol{x}}_n]$ by marginalizing out $\boldsymbol{\theta}$, $\boldsymbol{H}$, and $\boldsymbol{V}$ under a posterior over these variables, but this is analytically intractable. One alternative is to substitute maximum-a-posteriori (MAP) estimates of $\boldsymbol{\theta}$, $\boldsymbol{H}$, and $\boldsymbol{V}$ into Eqs. (15) and (16).

2.4. Fourier Trick

We here discuss the formulation of LD-PSDTF in the frequency domain. Using Eq. (14), the complex spectrum $\boldsymbol{F}\hat{\boldsymbol{x}}_n$ (linear transformation of $\hat{\boldsymbol{x}}_n$) is found to be complex-Gaussian distributed as follows:

$$\boldsymbol{F}\hat{\boldsymbol{x}}_n | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}} \sim \mathcal{N}_c\left(\mathbf{0}, \sum_{k=1}^K \hat{\theta}_k^2 \hat{h}_{kn}^2 \boldsymbol{F} \hat{\boldsymbol{V}}_k \boldsymbol{F}^H\right). \tag{17}$$

It is known that $\hat{\boldsymbol{V}}_k$ can be diagonalized by using $\boldsymbol{F}$ if $\hat{\boldsymbol{V}}_k$ is *strictly* a circulant matrix. A trivial example is a case that $\hat{\boldsymbol{V}}_k$ is a scaled identity matrix, *i.e.*, $\hat{\boldsymbol{w}}_{kn}$ is stationary white Gaussian noise. If $\hat{\boldsymbol{V}}_k$ is a periodic kernel and its size M is much larger than its period, $\hat{\boldsymbol{V}}_k$ can be *roughly* viewed as a circulant matrix.

These facts justify IS-NMF for power-spectrogram decomposition. Since music audio signals roughly consist of pitched sounds and percussive sounds, it is reasonable to approximate $\hat{\boldsymbol{V}}_k$ as a convex combination of

periodic kernels (for pitched sounds) and identity matrices (for percussive sounds). In the frequency domain LD-PSDTF thus reduces to IS-NMF discarding the covariance between frequency bins, while in the time domain the full covariance structure is still taken into account. This approximation dramatically reduces the computational cost of LD-PSDTF from $O(M^3NK)$ to $O(MNK)$ as suggested in (Liutkus et al., 2011).

3. Variational Inference

We explain an inference method for a Bayesian model of GaP-LD-PSDTF defined by Eqs. (4), (5), (6), and (8). Given the observed data $\boldsymbol{X}$, our goal is to calculate a posterior $p(\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V}|\boldsymbol{X})$ by using the Bayes rule $p(\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V}|\boldsymbol{X}) = p(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})/p(\boldsymbol{X})$. Since $p(\boldsymbol{X})$ is analytically intractable, we use a variational Bayesian (VB) method for approximating $p(\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V}|\boldsymbol{X})$ by a factorizable distribution $q(\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})$ given by

$$q(\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V}) = \prod_{k=1}^{K} \left(q(\theta_k) \left(\prod_{n=1}^{N} q(h_{kn}) \right) q(\boldsymbol{V}_k) \right). \quad (18)$$

These factors can be alternately updated to monotonically increase a *log-evidence lower bound* $\mathcal{L}$ given by

$$\begin{aligned} &\log p(\boldsymbol{X}) \geq \mathbb{E}[\log p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})] \\ &+ \mathbb{E}[\log p(\boldsymbol{\theta})] + \mathbb{E}[\log p(\boldsymbol{H})] + \mathbb{E}[\log p(\boldsymbol{V})] \\ &- \mathbb{E}[\log q(\boldsymbol{\theta})] - \mathbb{E}[\log q(\boldsymbol{H})] - \mathbb{E}[\log q(\boldsymbol{V})] \equiv \mathcal{L}. \quad (19) \end{aligned}$$

Since the first term is still intractable, we need to take a further lower bound $\mathcal{L}'$ such that $\mathcal{L} \geq \mathcal{L}'$. Note that $\mathcal{L}$ can be indirectly maximized by maximizing $\mathcal{L}'$. The updating formulas are

$$\begin{aligned} q(\boldsymbol{\theta}) &\propto p(\boldsymbol{\theta}) \exp(\mathbb{E}_{q(\boldsymbol{H},\boldsymbol{V})}[\log q(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})]), \\ q(\boldsymbol{H}) &\propto p(\boldsymbol{H}) \exp(\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{V})}[\log q(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})]), \quad (20) \\ q(\boldsymbol{V}) &\propto p(\boldsymbol{V}) \exp(\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{H})}[\log q(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})]), \end{aligned}$$

where $\log q(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})$ is a variational lower bound of $\log p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})$, which is given by Eq. (23).

3.1. Log-Evidence Lower Bound

To derive the tractable bound $\mathcal{L}'$, we focus on the convexity and concavity of matrix-variate functions over PSD matrices. For example, $f(\boldsymbol{V}) = \log|\boldsymbol{V}|$ is concave and $g(\boldsymbol{V}) = \mathrm{tr}(\boldsymbol{Z}\boldsymbol{V}^{-1})$ is convex for any PSD matrix $\boldsymbol{Z}$. Let M be the dimension of $\boldsymbol{V}$.

We first calculate a tangent plane of $f(\boldsymbol{V})$ by using a first-order Taylor expansion as follows:

$$\log|\boldsymbol{V}| \leq \log|\boldsymbol{\Omega}| + \mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{V}) - M, \quad (21)$$

where $\boldsymbol{\Omega}$ is an arbitrary PSD matrix (tangent point) and the equality is satisfied when $\boldsymbol{\Omega} = \boldsymbol{V}$.

We then use the following matrix inequality, proposed by Sawada et al. (2012), regarding $g(\boldsymbol{V})$:

$$\mathrm{tr}\left(\boldsymbol{Z} \left(\textstyle\sum_{k=1}^{K} \boldsymbol{V}_k \right)^{-1} \right) \leq \sum_{k=1}^{K} \mathrm{tr}\left(\boldsymbol{\Phi}_k \boldsymbol{Z} \boldsymbol{\Phi}_k^T \boldsymbol{V}_k^{-1} \right), \quad (22)$$

where $\{\boldsymbol{V}_k\}_{k=1}^K$ is a set of arbitrary PSD matrices, $\{\boldsymbol{\Phi}_k\}_{k=1}^K$ is a set of auxiliary matrices that sum to the identity matrix (*i.e.*, $\sum_k \boldsymbol{\Phi}_k = \boldsymbol{I}$), and the equality is satisfied when $\boldsymbol{\Phi}_k = \boldsymbol{V}_k(\sum_{k'} \boldsymbol{V}_{k'})^{-1}$.

Using Eqs. (21) and (22), we can derive the tractable lower bound of $\mathbb{E}[\log p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})]$ (the first term of Eq. (19)). A term regarding $\boldsymbol{X}_n$ is bounded as follows:

$$\begin{aligned} &\mathbb{E}[\log p(\boldsymbol{X}_n|\boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{V})] \quad (\text{see Eq. (7)}) \quad (23) \\ &= -\frac{\nu}{2}\mathbb{E}[\log|\boldsymbol{Y}_n|] - \frac{\nu}{2}\mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{X}_n \boldsymbol{Y}_n^{-1}\right)\right] + \text{const.} \\ &\geq -\frac{\nu}{2}\log|\boldsymbol{\Omega}_n| - \frac{\nu}{2}\textstyle\sum_k \mathbb{E}\left[\mathrm{tr}\left(\theta_k h_{kn} \boldsymbol{V}_k \boldsymbol{\Omega}_n^{-1}\right)\right] + \frac{\nu M}{2} \\ &\quad - \frac{\nu}{2}\textstyle\sum_k \mathbb{E}\left[\mathrm{tr}\left(\theta_k^{-1} h_{kn}^{-1} \boldsymbol{V}_k^{-1} \boldsymbol{\Phi}_{nk} \boldsymbol{X}_n \boldsymbol{\Phi}_{nk}^T\right)\right] + \text{const.}, \end{aligned}$$

where $\boldsymbol{\Omega}_n$ is a PSD matrix and $\{\boldsymbol{\Phi}_{nk}\}_{k=1}^K$ is a set of auxiliary matrices that sum to an identity matrix. Letting the partial derivatives of Eq. (23) equal to be zero, we can obtain the optimal values of $\boldsymbol{\Omega}_n$ and $\{\boldsymbol{\Phi}_{nk}\}_{k=1}^K$ that satisfy the equality as follows:

$$\boldsymbol{\Omega}_n = \textstyle\sum_k \mathbb{E}[\boldsymbol{Y}_{nk}], \quad (24)$$

$$\boldsymbol{\Phi}_{nk} = \left(\mathbb{E}\left[\boldsymbol{Y}_{nk}^{-1}\right]\right)^{-1} \left(\textstyle\sum_{k'} \left(\mathbb{E}\left[\boldsymbol{Y}_{nk'}^{-1}\right]\right)^{-1} \right)^{-1}. \quad (25)$$

3.2. Variational Bayesian Update

Here we discuss the functional forms of $q(\theta_k)$, $q(h_{kn})$, and $q(\boldsymbol{V}_k)$. A problem lies in the non-conjugacy of the Bayesian model. Eq. (23) involves the expectations both of the scalar variables and of their reciprocals. This is, the sufficient statistics are x and x^{-1}, although those of the gamma prior are $\log(x)$ and x. This means that the functional forms of $q(\theta_k)$ and $q(h_{kn})$ are given by the generalized inverse Gaussian (GIG) distribution, as shown in Hoffman et al. (2010). Note that the GIG distribution is defined as

$$\mathrm{GIG}(x|\gamma, \rho, \tau) = \frac{(\rho/\tau)^{\frac{\gamma}{2}}}{2K_\gamma(\sqrt{\rho\tau})} x^{\gamma-1} e^{-\frac{1}{2}(\rho x + \tau x^{-1})}, \quad (26)$$

where γ, $\rho > 0$, and $\tau > 0$ are parameters and K_γ is the modified Bessel function of the second kind. The expectations $\mathbb{E}[x]$ and $\mathbb{E}[x^{-1}]$ are given by

$$\mathbb{E}[x] = \frac{\sqrt{\tau}K_{\gamma+1}(\sqrt{\rho\tau})}{\sqrt{\rho}K_\gamma(\sqrt{\rho\tau})}, \ \mathbb{E}\left[\frac{1}{x}\right] = \frac{\sqrt{\rho}K_{\gamma-1}(\sqrt{\rho\tau})}{\sqrt{\tau}K_\gamma(\sqrt{\rho\tau})}. \quad (27)$$

As to matrix variable $\boldsymbol{V}_k$, we found that the functional form of $q(\boldsymbol{V}_k)$ is given by the matrix GIG (MGIG) dis-

tribution (Barndorff-Nielsen et al., 1982). The MGIG distribution over PSD matrix $\boldsymbol{X}$ is defined as

$$\mathrm{MGIG}(\boldsymbol{X}|\gamma,\boldsymbol{R},\boldsymbol{T}) = \frac{2^{\gamma M}}{|\boldsymbol{T}|^{\gamma} B_{\gamma}(\boldsymbol{R}\boldsymbol{T}/4)} |\boldsymbol{X}|^{\gamma-\frac{M+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{R}\boldsymbol{X}+\boldsymbol{T}\boldsymbol{X}^{-1}\right)\right), \quad (28)$$

where γ is a real number, $\boldsymbol{R},\boldsymbol{T}>\boldsymbol{0}$ are PSD matrices, M is the size of $\boldsymbol{X}$, and B_{γ} is the matrix Bessel function of the second kind (Herz, 1955). It includes the Wishart distribution as a special case (Butler, 1998) and its sufficient statistics are $\log|\boldsymbol{X}|$, $\boldsymbol{X}$, and $\boldsymbol{X}^{-1}$. To calculate $\mathbb{E}[\boldsymbol{X}]$ and $\mathbb{E}[\boldsymbol{X}^{-1}]$, we use a Monte Carlo method as described in the supplementary material.

Consequently, we can assume the following forms:

$$\begin{aligned} q(\theta_k) &= \mathrm{GIG}(\theta_k|\gamma_k^{\theta},\rho_k^{\theta},\tau_k^{\theta}),\\ q(h_{kn}) &= \mathrm{GIG}(h_{kn}|\gamma_{kn}^{h},\rho_{kn}^{h},\tau_{kn}^{h}),\\ q(\boldsymbol{V}_k) &= \mathrm{MGIG}(\boldsymbol{V}_k|\gamma_k^{V},\boldsymbol{R}_k^{V},\boldsymbol{T}_k^{V}). \end{aligned} \quad (29)$$

These parameters are iteratively updated as follows:

$$\begin{aligned} &\gamma_k^{\theta} = \alpha c/K, \quad \rho_k^{\theta} = 2\alpha + \nu\textstyle\sum_n \mathrm{tr}\left(\mathbb{E}[h_{kn}\boldsymbol{V}_k]\,\boldsymbol{\Omega}_n^{-1}\right),\\ &\tau_k^{\theta} = \nu\textstyle\sum_n \mathrm{tr}\left(\mathbb{E}\left[h_{kn}^{-1}\boldsymbol{V}_k^{-1}\right]\boldsymbol{\Phi}_{nk}\boldsymbol{X}_n\boldsymbol{\Phi}_{nk}^T\right),\\ &\gamma_{kn}^{h} = a_0, \quad \rho_{kn}^{h} = 2b_0 + \nu\mathrm{tr}\left(\mathbb{E}[\theta_k\boldsymbol{V}_k]\,\boldsymbol{\Omega}_n^{-1}\right),\\ &\tau_{kn}^{h} = \nu\mathrm{tr}\left(\mathbb{E}\left[\theta_k^{-1}\boldsymbol{V}_k^{-1}\right]\boldsymbol{\Phi}_{nk}\boldsymbol{X}_n\boldsymbol{\Phi}_{nk}^T\right),\\ &\gamma_k^{V} = \nu_0/2, \quad \boldsymbol{R}_k^{V} = \boldsymbol{V}_0^{-1} + \nu\textstyle\sum_n \mathbb{E}[\theta_k h_{kn}]\,\boldsymbol{\Omega}_n^{-1},\\ &\boldsymbol{T}_k^{V} = \nu\textstyle\sum_n \mathbb{E}\left[\theta_k^{-1}h_{kn}^{-1}\right]\boldsymbol{\Phi}_{nk}\boldsymbol{X}_n\boldsymbol{\Phi}_{nk}^T. \end{aligned} \quad (30)$$

3.3. Multiplicative Update

The multiplicative update (MU) is a well-known optimization technique often used for maximum-likelihood estimation of NMF. To show a clear connection of LD-PSDTF to IS-NMF, we derive closed-form MU rules for calculating the point estimates of $\boldsymbol{H}$ and $\boldsymbol{V}$. Note that we assume $\theta_k = 1$ and $\mathrm{tr}(\boldsymbol{V}_k) = 1$ (unit trace) to remove the scale arbitrariness. If $\mathrm{tr}(\boldsymbol{V}_k) = s$, the scale adjustments $\boldsymbol{V}_k \leftarrow \frac{1}{s}\boldsymbol{V}_k$ and $\boldsymbol{h}_k \leftarrow s\boldsymbol{h}_k$ do not change the LD divergence $\mathcal{D}_{\mathrm{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$.

We aim to maximize the log-likelihood given by removing the expectation operators from Eq. (23). Letting the partial derivative with respect to h_{kn} equal to be zero, we get the following update rule:

$$h_{kn} \leftarrow h_{kn}\sqrt{\frac{\mathrm{tr}\left(\boldsymbol{Y}_n^{-1}\boldsymbol{V}_k\boldsymbol{Y}_n^{-1}\boldsymbol{X}_n\right)}{\mathrm{tr}\left(\boldsymbol{Y}_n^{-1}\boldsymbol{V}_k\right)}}. \quad (31)$$

Then, letting the partial derivative with respect to $\boldsymbol{V}_k$ equal to be zero, we get the following equation:

$$\boldsymbol{V}_k\boldsymbol{P}_k\boldsymbol{V}_k = \boldsymbol{V}_k^{\mathrm{old}}\boldsymbol{Q}_k\boldsymbol{V}_k^{\mathrm{old}}, \quad (32)$$

where $\boldsymbol{P}_k$ and $\boldsymbol{Q}_k$ are PSD matrices given by

$$\boldsymbol{P}_k = \sum_{n=1}^{N} h_{kn}\boldsymbol{Y}_n^{-1}, \ \boldsymbol{Q}_k = \sum_{n=1}^{N} h_{kn}\boldsymbol{Y}_n^{-1}\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}. \quad (33)$$

Sawada et al. (2012) derived a complicated solution of Eq. (32), but we can solve it analytically by using the Cholesky decomposition $\boldsymbol{Q}_k = \boldsymbol{L}_k\boldsymbol{L}_k^T$, where $\boldsymbol{L}_k$ is a lower triangular matrix. Finally, we get

$$\boldsymbol{V}_k \leftarrow \boldsymbol{V}_k\boldsymbol{L}_k(\boldsymbol{L}_k^T\boldsymbol{V}_k\boldsymbol{P}_k\boldsymbol{V}_k\boldsymbol{L}_k)^{-\frac{1}{2}}\boldsymbol{L}_k^T\boldsymbol{V}_k. \quad (34)$$

When all the matrices are diagonal, Eqs. (31) and (34) reduce to the one in IS-NMF (Nakano et al., 2010).

4. Related Work

We show that PSDTF has deep connections to nonnegative matrix factorization (NMF), tensor factorization (TF), and principal component analysis (PCA).

4.1. Nonnegative Matrix Factorization

PSDTF includes NMF as a special case. If we restrict PSD matrices $\boldsymbol{X}_n$ and $\boldsymbol{V}_k$ to diagonal matrices (*i.e.*, $\boldsymbol{X}_n = \mathrm{diag}(\boldsymbol{x}_n)$ and $\boldsymbol{V}_k = \mathrm{diag}(\boldsymbol{v}_k)$ for some *nonnegative* vectors $\boldsymbol{x}_n$ and $\boldsymbol{v}_k$) Eq. (1) can be written as

$$\boldsymbol{x}_n \approx \sum_{k=1}^{K} \theta_k h_{kn}\boldsymbol{v}_k, \quad (35)$$

where θ_k and h_{kn} are *nonnegative* numbers. If $\theta_k = 1$, this model reduces to the basic formulation of NMF. Févotte et al. (2009) showed that the IS divergence is theoretically suitable for evaluating the reconstruction error of Eq. (35) in the task of audio source separation. Hoffman et al. (2010) proposed an infinite extension of IS-NMF (GaP-IS-NMF) using a gamma process prior on $\boldsymbol{\theta}$. GaP-LD-PSDTF can therefore be viewed as a natural extension of GaP-IS-NMF.

An interesting finding is that the *positive semidefiniteness* (nonnegative definiteness) constraint on matrices in PSDTF induces sparse decomposition like the *nonnegativity* constraint on vectors and scalars as in NMF. Positive semidefiniteness can therefore be considered a generalization of the nonnegativity concept.

4.2. Tensor Factorization

PSDTF is related to a variant of TF called canonical polyadic (CP) decomposition (Carroll & Chang, 1970; Harshman, 1970). If we restrict a PSD matrix $\boldsymbol{V}_k$ to a rank-1 matrix (*i.e.*, $\boldsymbol{V}_k = \boldsymbol{u}_k\boldsymbol{u}_k^T$ for some vector $\boldsymbol{u}_k$), Eq. (1) can be written as

$$\boldsymbol{X} \approx \sum_{k=1}^{K} \theta_k\boldsymbol{h}_k \otimes \boldsymbol{u}_k \otimes \boldsymbol{u}_k. \quad (36)$$

This can be viewed as CP decomposition in which basis vectors of the second mode, $\{\boldsymbol{u}_k\}_{k=1}^K$, are constrained to be equal to those of the third mode, and θ_k and $\boldsymbol{h}_k$ should be nonnegative. In addition, PSDTF uses the LD divergence for evaluating the reconstruction error while typical TF uses the Euclidean distance.

There are some other related models. Tucker decomposition (Tucker, 1966) is a generalization of CP decomposition and Xu et al. (2012) proposed its infinite extension ($K \to \infty$) based on the Gaussian or t process. Shashua & Hazan (2005) proposed nonnegative TF (NTF) that, like NMF, imposes a nonnegativity constraint on all elements of factors. In this paper the nonnegativity constraint on θ_k and $\boldsymbol{h}_k$, not on $\boldsymbol{u}_k$, led to a new class of TF.

4.3. Principal Component Analysis

LD-PSDTF is related to a major class of matrix factorization (MF) using the Gaussian distribution as a core building block of probabilistic models. Let us recall the MF model given by Eq. (11):

$$\hat{\boldsymbol{X}} = \hat{\boldsymbol{W}}\hat{\boldsymbol{H}}^T \tag{37}$$

where $\hat{\boldsymbol{X}} \in \mathbb{R}^{M\times N}$, $\hat{\boldsymbol{W}} \in \mathbb{R}^{M\times K}$, and $\hat{\boldsymbol{H}} \in \mathbb{R}^{N\times K}$ (N is the number of observations).

Several probabilistic models are obtained by putting Gaussian priors in different ways. Placing an isotropic Gaussian prior on N columns of $\hat{\boldsymbol{H}}$ (latent-space coordinates corresponding to the observations) leads to probabilistic PCA (PPCA) (Bishop, 1999). If we put an isotropic Gaussian prior on M rows of $\hat{\boldsymbol{W}}$ (mapping functions from the latent space to the observed space), the resulting model is called dual PPCA. Marginalizing $\hat{\boldsymbol{W}}$ out, we can formulate a Gaussian process latent variable model (GPLVM) (Lawrence, 2003). PSDTF, on the other hand, puts a *full-covariance* Gaussian prior on K columns of $\hat{\boldsymbol{W}}$. If we instead use a GP prior, PSDTF given by Eq. (14) can be viewed as multiple kernel learning (MKL) (Lanckriet et al., 2004).

5. Evaluation

This section reports experiments to evaluate the performance of LD-PSDTF.

5.1. Synthetic Data

We evaluated the capability of GaP-LD-PSDTF to discover basis PSD matrices $\{\boldsymbol{V}_k\}_{k=1}^{K^+}$ used for generating an observed tensor $\boldsymbol{X}$ and to estimate the value of K^+. In this experiment we use $M = \nu = 10$, $N = 2000$, and $K^+ = 6$. The synthetic data $\boldsymbol{X}$ was stochastically generated according to the following process:

$$\begin{aligned} h_{kn} &\sim \mathcal{G}(0.1, 0.1), \\ \boldsymbol{V}_k &\sim \mathcal{W}(10, \boldsymbol{I}/10), \\ \nu\boldsymbol{X}_n &\sim \mathcal{W}\left(\nu, \textstyle\sum_k h_{kn}\boldsymbol{V}_k\right), \end{aligned} \tag{38}$$

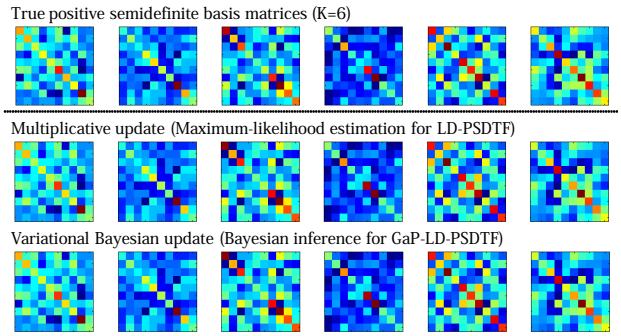

Figure 3. Experimental results for synthetic data.

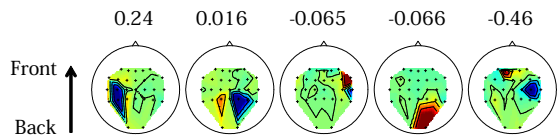

Figure 4. Brain activities discovered from EEG data.

To learn the GaP-LD-PSDTF model, we used the VB algorithm with a truncation level $K = 100$, and hyperparameters $\alpha = c = 1$, $a_0 = b_0 = 0.1$, $\nu_0 = 10$, and $\boldsymbol{V}_0 = \boldsymbol{I}/\nu_0$. Since Monte Carlo simulation of $\mathbb{E}[\boldsymbol{V}_k]$ and $\mathbb{E}[\boldsymbol{V}_k^{-1}]$ was found to be often unreliable, we instead calculated the maximum-a-posteriori estimates. For maximum-likelihood estimation of the LD-PSDTF model, we used the MU algorithm with $K = 6$. The both methods were initialized randomly.

As shown in Figure 3, the experimental results showed that the both models successfully discovered the correct basis matrices. The hyperparameters were not sensitive to the results. In GaP-LD-PSDTF, the true number of bases, $K^+ = 6$, was correctly estimated.

5.2. EEG Data

We then tested LD-PSDTF on a popular EEG dataset (Blankertz, 2001). We aimed to predict a left or right hand movement (label -1 or 1) from 500 ms EEG signals recorded at 28 channels of the brain ($M = 28$) with a sampling rate of 100 Hz. There were 416 trials ($N = 416$) of which 100 belong to the test set. For each trial we calculated a full-rank covariance matrix over 50 frames. The PSD basis matrices and their activations were estimated by using the MU algorithm with $K = 5$ in an unsupervised manner. We used Fisher's LDA for binary classification of the K-dimensional activations corresponding to the individual trials.

The most significant principal component of each basis matrix is shown in Figure 4, in which each number indicates the correlation between the ground-truth labels and the estimated activations on the test set. The accuracies of classification were 73% ($K = 5$) and 79% ($K = 10$). The leftmost and rightmost bases with high correlations are compatible with the well-known physiological process called event-related desynchronization (ERD). We consider these results promising.

Note that the best results from the competition were obtained by combining the first-order features (lateralized readiness potential: LRP) and the second-order ones (ERD), whereas our method used only the latter. Finding discriminative patterns without any supervision in this context is a highly nontrivial task, because discriminative signals are much weaker than irrelevant oscillatory activities (*e.g.*, occipital alpha waves).

5.3. Music Data

We evaluated LD-PSDTF for music signal analysis. As discussed in Section 2.4, LD-PSDTF formulated in the time domain is equivalent to IS-NMF formulated in the frequency domain if $\{\boldsymbol{V}_k\}_{k=1}^K$ are circulant matrices such as periodic kernels, identity matrices, and their conic sums. Since this is a reasonable assumption for music signals, we used the Fourier trick explained in Section 2.4. We tested an infinite model with a truncation level $K = 100$ and finite models ($\theta_k = 1$) with different K ranging from 1 to 100. For comparison, we tested an infinite model of KL-NMF (GaP-KL-NMF) and finite models with different K ranging from 1 to 100 for amplitude-spectrogram decomposition.

We used three songs (No.1, 2, and 3) from the "RWC Music Database: Popular Music" (Goto et al., 2002). The CD-quality audio signals were downsampled at 16 [kHz] and were analyzed by using short-time Fourier transform with a window size of 128 [ms] and a shifting interval of 64 [ms]. The size of $\boldsymbol{X}$ is specified as $M = 2048$, and $N = 3237$, 3447, and 3020, respectively. We used the VB algorithm with hyperparameters $\alpha = 1$, $c = 1$, $a_0 = b_0 = 0.1$, $\nu_0 = M$, $\nu = 1$, and $\boldsymbol{V}_0 = \boldsymbol{I}/\nu_0$.

The experimental results showed that in each song the value of K^+ chosen by GaP-LD-PSDTF was close to the best value of K found by finite models (Figure 5). This proves that GaP-LD-PSDTF has an ability of automatic model-order selection without expensive grid search. Similar results were obtained in KL-NMF, but GaP-LD-PSDTF achieved much higher log-evidence lower bounds than GaP-KL-NMF did. This supports the appropriateness of IS-NMF for music analysis. Additional results of source separation with sound samples are given in the supplementary material.

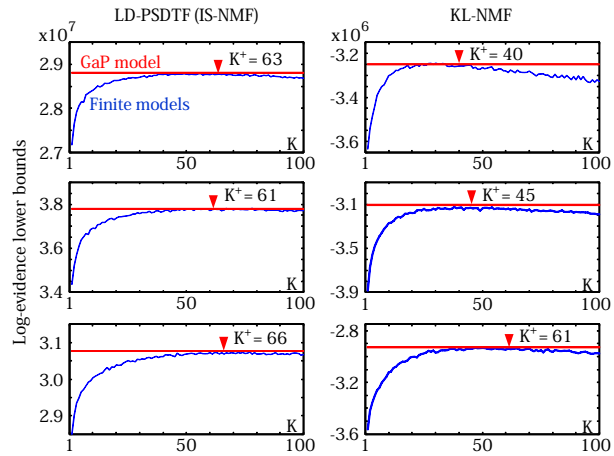

Figure 5. Experimental results for music recordings.

6. Conclusion

This paper presented positive semidefinite tensor factorization (PSDTF) as a natural extension of nonnegative matrix factorization (NMF). We used a Bregman matrix divergence called the log-determinant (LD) divergence as the reconstruction error. This LD-PSDTF can be viewed as a natural extension of NMF based on the Itakura-Saito (IS) divergence. We formulated a nonparametric Bayesian model that allows an observed tensor to contain an unbounded number of bases and derived a variational Bayesian algorithm and a multiplicative update rule by using matrix inequalities. In addition, we showed the effectiveness of the Fourier trick, *i.e.,* the frequency-domain formulation can dramatically reduce the computational cost in some applications such as music signal analysis.

One interesting open question is what kind of PSDTF can be viewed as an extension of NMF based on the Kullback-Leibler (KL) divergence. The von-Neumann (vN) divergence (Tsuda et al., 2005) is well known as another major type of the Bregman matrix divergence that includes the KL divergence as a special case. Substituting $\phi(\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{Z}\log\boldsymbol{Z} - \boldsymbol{Z})$ into Eq. (2), the vN divergence is given by

$$\mathcal{D}_{\mathrm{vN}}(\boldsymbol{X}_n|\boldsymbol{Y}_n) = \mathrm{tr}\left(\boldsymbol{X}_n\log\boldsymbol{X}_n - \boldsymbol{X}_n\log\boldsymbol{Y}_n - \boldsymbol{X}_n + \boldsymbol{Y}_n\right). \quad (39)$$

The assumption underlying $p(\boldsymbol{X}_n|\boldsymbol{Y}_n)$, however, is not obvious although the Bregman divergence must correspond one-to-one to an exponential family. We plan to investigate vN-PSDTF to formulate a Bayesian model.

Acknowledgment: This study was partially supported by JSPS KAKENHI 23700184, MEXT KAKENHI 25870192, and JST OngaCREST project. We thank the reviewers for giving insightful comments.

References

Barndorff-Nielsen, O., Blæsild, P., Jensen, J. L., and Jørgensen, B. Exponential transformation models. *Royal Society of London*, 379(1776):41–65, 1982.

Bishop, C. M. Variational principal components. In *ICANN*, pp. 509–514, 1999.

Blankertz, B., Curio, G., and Müller, K.-L. Classifying single trial EEG: Towards brain computer interfacing, In: *NIPS*, 2001. www.bbci.de/competition/ii/berlin_desc.html

Bregman, L. M. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR CMMP*, 7(3):200–217, 1967.

Butler, R. W. Generalized inverse Gaussian distributions and their Wishart connections. *Scandinavian Journal of Statistics*, 25(1):69–75, 1998.

Butler, R. W. and Wood, A. Laplace approximation for Bessel functions of matrix argument. *J. of Computational and Applied Math.*, 155(2):359–382, 2003.

Carroll, J. D. and Chang, J. J. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, 35(3):283–319, 1970.

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

Févotte, C., Bertin, N., and Durrieu, J.-L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

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

Harshman, R. A. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16(1), 1970.

Herz, C. S. Bessel functions of matrix argument. *Annals of Mathematics*, 61(3):474–523, 1955.

Hoffman, M., Blei, D., and Cook, P. Bayesian nonparametric matrix factorization for recorded music. In *ICML*, pp. 439–446, 2010.

Itakura, F. and Saito, S. Analysis synthesis telephony based on the maximum likelihood method. In *ICA*, pp. C17–C20, 1968.

Kulis, B., Sustik, M., and Dhillon, I. Low-rank kernel learning with Bregman matrix divergences. *JMLR*, 10:341–376, 2009.

Kullback, S. and Leibler, R. On information and sufficiency. *Annals of Math. Stat.*, 22(1):79–86, 1951.

Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., and Jordan, M. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004.

Lawrence, N. D. Gaussian process latent variable models for visualisation of high dimensional data. In *NIPS*, 2003.

Lee, D. and Seung, H. Algorithms for non-negative matrix factorization. In *NIPS*, pp. 556–562, 2000.

Lee, H., Cichocki, A., and Choi, S. Nonnegative matrix factorization for motor imagery EEG classification. In *ICANN*, pp. 250–259, 2006.

Liutkus, A., Badeau, R., and Richard, G. Gaussian processes for underdetermined source separation. *IEEE Trans. on ASLP*, 59(7):3155–3167, 2011.

Nakano, M., Kameoka, H., Roux, J. Le, Kitano, Y., Ono, N., and Sagayama, S. Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta divergence. In *MLSP*, pp. 283–288, 2010.

Salakhutdinov, R. and Mnih, A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *ICML*, pp. 880–887, 2008.

Sawada, H., Kameoka, H., Araki, S., and Ueda, N. Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization. In *ICASSP*, pp. 261–264, 2012.

Shashua, A. and Hazan, T. Non-negative tensor factorization with applications to statistics and computer vision. In *ICML*, pp. 792–799, 2005.

Sivalingam, R., Boley, D., Morellas, V., and Papanikolopoulos, N. Tensor sparse coding for region covariances. In *ECCV*, pp. 722–735, 2010.

Smaragdis, P. and Brown, J. C. Non-negative matrix factorization for polyphonic music transcription. In *WASPAA*, pp. 177–180, 2003.

Tsuda, K., Rätsch, G., and Warmuth, M. K. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *JMLR*, 6:995–1018, 2005.

Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

Xu, Z., Yan, F., and Qi, Y. Infinite Tucker decomposition: Nonparametric Bayesian models for multiway data analysis. In *ICML*, pp. 1023–1030, 2012.