Chapter 2
Non-negative Matrix Factorization and Its Variants for Audio Signal Processing

Hirokazu Kameoka

Abstract In this chapter, I briefly introduce a multivariate analysis technique called non-negative matrix factorization (NMF), which has attracted a lot of attention in the field of audio signal processing in recent years. I will mention some basic properties of NMF, effects induced by the non-negative constraints, how to derive an iterative algorithm for NMF, and some attempts that have been made to apply NMF to audio processing problems.

Keywords Non-negative matrix factorization · Majorization-minimization algorithm · Bregman divergence · Bayesian nonparametrics · Audio signal processing

2.1 Introduction

There are many kinds of real-world data given by non-negative values, such as power spectra, pixel values and count data. In a way similar to multivariate analysis techniques such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), decomposing non-negative data into the sum of the underlying components can be useful in many situations: For example, if we can extract the power spectra of the underlying sources in a mixture signal, they may be useful for noise reduction and source separation. If we can decompose face images into components corresponding to facial features such as the eyes, nose and mouth, they may be useful for face recognition, identification and synthesis. If we can decompose the word histograms of text documents into components associated with latent topics such as politics, sport and economy, they may be useful for document indexing and retrieval. Similarly, if we can extract patterns reflecting users' preferences from purchase logs,

H. Kameoka (✉)
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail: kameoka@hil.t.u-tokyo.ac.jp; kameoka.hirokazu@lab.ntt.co.jp

H. Kameoka
Nippon Telegraph and Telephone Corporation, 3-1 Morinosato Wakamiya, Atsugi,
Kanagawa 243-0198, Japan

T. Sakata (ed.), *Applied Matrix and Tensor Variate Data Analysis*,
JSS Research Series in Statistics, DOI 10.1007/978-4-431-55387-8_2

they may be useful for making recommendations. A multivariate analysis technique enabling the decomposition of non-negative data into non-negative components is called Non-negative Matrix Factorization (NMF) [1]. In this chapter, I will mention some basic properties of NMF, how to derive an iterative algorithm for NMF, and some attempts that have been made to apply NMF and its variants to audio processing problems.

2.2 What Is NMF?

In the following, we will represent data by vectors. For image data, each pixel value will correspond to a single element of the data vector. For power spectrum data, the power at each frequency point will correspond to a single element of the data vector. Let us assume that we are given a set of N non-negative data vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \in \mathbb{R}^{\geq 0, K}$. We refer to each of them as an observed vector. Here, $\mathbb{R}^{\geq 0, K}$ is used to represent an entire set of K-dimensional non-negative vectors. The aim of NMF is to decompose each of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ into the sum of M non-negative components: The problem is to find the linear combinations of M basis vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M \in \mathbb{R}^{\geq 0, K}$ that best approximate $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$:

$$\boldsymbol{y}_n \simeq \sum_{m=1}^{M} \boldsymbol{h}_m u_{m,n} \quad (n = 1, \ldots, N), \tag{2.1}$$

subject to the non-negativity constraints on both the basis vectors $\boldsymbol{h}_m$ and the coefficients $u_{m,n}$. Here, it is important to note that the observed data are assumed to be quantities that are additive in nature. Although neither a pixel value nor a power spectrum is strictly an additive quantity, we must be aware of the fact that when applying NMF, the additivity of the data of interest will be implicitly assumed to hold, regardless of whether this assumption is true or only approximately true. The non-additivity of power spectra will be discussed in detail in Sect. 2.7. In addition to the additivity assumption as regards the data, the non-negativity constraint is one of the most important features of NMF. As explained later, the non-negativity constraint contributes to inducing sparsity of both the basis vectors and the coefficients.

Now, if we let $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N] = (y_{k,n})_{K \times N}$, $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M] = (h_{k,m})_{K \times M}$ and $\boldsymbol{U} = (u_{m,n})_{M \times N}$, Eq. (2.1) can be rewritten as $\boldsymbol{Y} \simeq \boldsymbol{H}\boldsymbol{U}$. NMF can thus be seen as a problem of factorizing an observed matrix into the product of two non-negative matrices, which gives NMF its name. To understand NMF intuitively, see Fig. 2.1 for an example of NMF applied to the spectrogram of an audio signal, interpreted as a non-negative matrix.

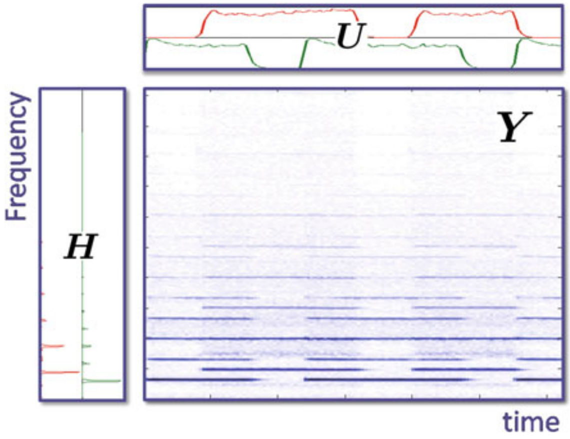

Fig. 2.1 NMF applied to the spectrogram of an audio signal. Each column of $\boldsymbol{H}$ and each row of $\boldsymbol{U}$ can be interpreted as a spectral template and the corresponding temporal activation, respectively. $\boldsymbol{HU}$ can thus be viewed as the sum of the spectral templates scaled by time-varying amplitudes

2.3 Basic Properties of NMF

The number M of basis vectors is usually set smaller than the dimension K and the number N of data vectors. This is because when $M \geq K$ or $M \geq N$, there are trivial solutions to the factorization $\boldsymbol{Y} = \boldsymbol{HU}$. For example, when $M = K$, we have $\boldsymbol{Y} = \boldsymbol{IU}$ and when $M = N$, we have $\boldsymbol{Y} = \boldsymbol{HI}$, where $\boldsymbol{I}$ denotes an identity matrix. Obviously, neither of these decompositions provides information about the latent components underlying the data. When $M < \min(K, N)$, the factorization amounts to approximating the data matrix using a lower rank matrix, which provides meaningful information about the latent components. Geometrically, while PCA (singular value decomposition) tries to find a linear subspace to which observed vectors belong, NMF can be interpreted as finding a convex cone (see Fig. 2.2) that is closest to the entire set of observed vectors. The number M of basis vectors corresponds to the dimension of the convex cone, which depends on the data and is usually unknown. Thus, determining M is an important issue in NMF. Recent techniques for determining M will be mentioned in Sect. 2.8.

With NMF, the elements of the coefficient matrix $\boldsymbol{U}$ tend to become sparse as a side effect of the non-negativity constraint. The intuitive reason for this can be explained as follows. First, let us consider an unconstrained optimization problem $\hat{\boldsymbol{u}} = \underset{\boldsymbol{u}}{\operatorname{argmin}}\, \mathcal{D}(\boldsymbol{y}|\boldsymbol{Hu})$ where $\mathcal{D}(\cdot|\cdot)$ is a measure of the difference between two vectors. $\boldsymbol{H}\hat{\boldsymbol{u}}$ corresponds to the closest point from $\boldsymbol{y}$ in the subspace spanned by $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M$. If $\mathcal{D}$ is defined as an ℓ_2 norm, for example, this point simply corresponds to the orthogonal projection of $\boldsymbol{y}$ onto the subspace. Now, let us denote the solution to this optimization problem under the non-negativity constraint by $\tilde{\boldsymbol{u}}$. Except for a coincidental case where the unconstrained optimal solution $\hat{\boldsymbol{u}}$ satisfies the non-negativity constraint, $\boldsymbol{H}\tilde{\boldsymbol{u}}$ will be a closest point to $\hat{\boldsymbol{u}}$ in the convex cone shown in Fig. 2.2, namely some point on the boundary of the cone. This means at least one of the elements of the coefficient vector becomes 0. Therefore, the con-

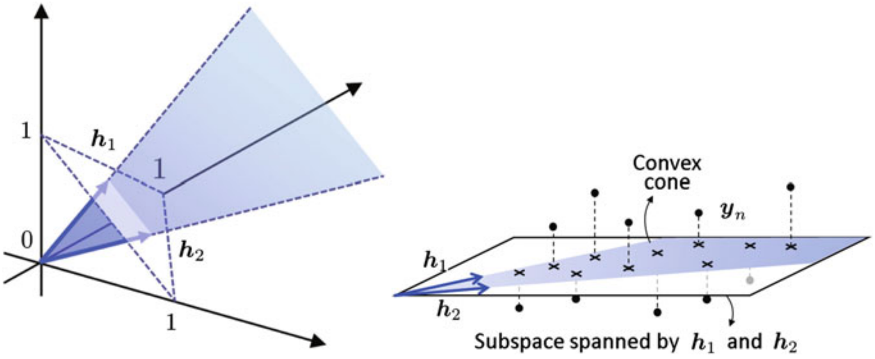

Fig. 2.2 Geometric understanding of NMF. Because of the non-negativity of $\boldsymbol{H}$, all basis vectors lie in the first quadrant. Because of the non-negativity of $\boldsymbol{U}$, $\boldsymbol{H}\boldsymbol{u}_n$ can only cover the area enclosed by the extended *lines* of all the basis vectors. Thus, NMF can be interpreted as finding a convex cone that is closest to the entire set of observed vectors

strained optimal solution $\tilde{\boldsymbol{u}}$ becomes relatively sparser (with a larger number of zero entries) than the unconstrained optimal solution $\hat{\boldsymbol{u}}$. This explains why NMF tends to produce sparse representations. It is important to note that sparsity is related to statistical independence (non-Gaussianity). Thus, roughly speaking, producing sparse representations implies that each row of the coefficient matrix tends to become uncorrelated. The above property also applies to the transposition of $\boldsymbol{Y} \simeq \boldsymbol{H}\boldsymbol{U}$, i.e., $\boldsymbol{Y}^{\mathsf{T}} \simeq \boldsymbol{U}^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}}$, meaning that $\boldsymbol{H}$ also tends to become sparse owing to the non-negativity constraint on $\boldsymbol{H}$.

2.4 NMF Algorithms

2.4.1 Positive Matrix Factorization and NMF

The original concept of NMF was first introduced by Paatero and Tapper in 1994 [2]. Within their formulation, they used the Frobenius norm of $\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{U}$ as a measure of the difference between $\boldsymbol{Y}$ and $\boldsymbol{H}\boldsymbol{U}$ and a logarithmic barrier function

$$B(\boldsymbol{H}, \boldsymbol{U}) = -\sum_{k,m} \log h_{k,m} - \sum_{m,n} \log u_{m,n} \tag{2.2}$$

as a penalizing term for violations of the non-negativity constraint, which approaches infinity as $h_{k,m}$ or $u_{m,n}$ approaches zero. They proposed a gradient-based optimization algorithm for minimizing the cost function defined as a weighted sum of these two terms.

Because of the property of the logarithmic barrier function, the elements of the matrices given by this method must always be positive (they never become zero). Thus, it is usually called “Positive Matrix Factorization (PMF)”, which distinguishes it from NMF. Several years later, Lee and Seung proposed an iterative scheme called the multiplicative update algorithm, which ensures the non-negativity of $\boldsymbol{H}$ and $\boldsymbol{U}$ without using barrier functions [1]. Owing to the simplicity of its implementation, NMF has subsequently gained considerable momentum in a wide range of research areas.

2.4.2 Divergence Measures

NMF leads to different optimization problems according to the definition of the measure of the difference between $\boldsymbol{Y}$ and $\boldsymbol{H}\boldsymbol{U}$. Lee and Seung have proposed deriving NMF algorithms using the Frobenius norm and the generalized Kullback-Leibler (KL) divergence (also known as the I divergence) [3] as the goodness-of-fit criteria. Of course, the optimal values of $\boldsymbol{H}$ and $\boldsymbol{U}$ depend on the choice of these criteria. It is desirable that the goodness-of-fit criterion be set according to the underlying generative process of the data $\boldsymbol{Y}$. For example, the Itakura-Saito (IS) divergence is often used as the model-fitting criterion for NMF when it is applied to power spectrum data [4, 5]. This is actually based on an assumption about the generative process of time-domain signals (as explained in Sect. 2.7.3).

For $y, x \in \mathbb{R}$, the Euclidean distance (squared error), the generalized KL divergence and the IS divergence of x from y are defined as

$$\mathcal{D}_{\mathrm{EU}}(y|x) = (y-x)^2, \tag{2.3}$$

$$\mathcal{D}_{\mathrm{KL}}(y|x) = y \log \frac{y}{x} - y + x, \tag{2.4}$$

$$\mathcal{D}_{\mathrm{IS}}(y|x) = \frac{y}{x} - \log \frac{y}{x} - 1, \tag{2.5}$$

respectively. All of these metrics become 0 only when $x = y$ and increase monotonically as x and y become more distant. Figure 2.3 shows the graph of each of these measures seen as a function of x. While the Euclidean distance is symmetric about $x = y$, the generalized KL divergence and the IS divergence are asymmetric and impose larger penalties when x is below y than when x is above y. It is also important to note that the IS divergence is invariant under the scaling of x and y since it is represented using only the ratio of x to y. By using these metrics, we can measure the difference between $\boldsymbol{H}\boldsymbol{U}$ and $\boldsymbol{Y}$ with

$$D_{\cdot}(\boldsymbol{H}, \boldsymbol{U}) = \sum_{k,n} \mathcal{D}_{\cdot}\Big(y_{k,n} \Big| \sum_m h_{k,m} u_{m,n}\Big),$$

where $\cdot$ indicates EU, KL or IS.

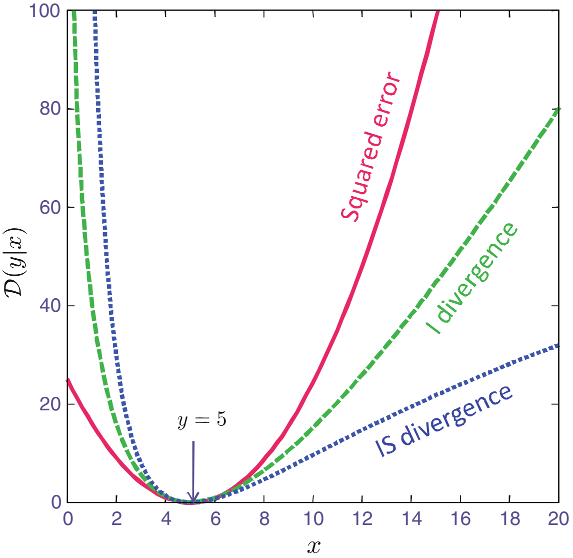

Fig. 2.3 Graph of $\mathcal{D}_{\mathrm{EU/KL/IS}}(y|x)$ as a function of x

2.4.3 Auxiliary Function Approach

The goal of NMF is to find optimal values for $\boldsymbol{H}$ and $\boldsymbol{U}$ that minimize one of these kinds of measures subject to the non-negativity constraint. Although it is usually difficult to obtain an analytical expression of the global optimum solution, one of the local optimum solutions can be searched for numerically using the "auxiliary function approach" (also known as the "Majorization-Minimization algorithm") [7, 25]. As explained later, the auxiliary function approach makes it possible to locally minimize an objective function by iteratively minimizing an auxiliary function whose lower bound is exactly equal to the objective function value. It should be noted that the Expectation-Maximization (EM) algorithm [8], a popular technique for maximum likelihood estimation from incomplete data, is a special case of this approach.

In NMF, the non-negativity constraint must be considered. If the objective function were given as the sum of individual terms, each relating to one matrix element, solving the constrained optimization problem would be relatively simple. But of course this is not the case. If we can use such a function as an auxiliary function, the constrained optimization problem of NMF can be solved in an iterative manner using the auxiliary function approach.

The definition of the auxiliary function and the principle of the auxiliary function approach are as follows.

Definition 2.1 Given an objective function $D(\theta)$ with the parameter $\theta = \{\theta_i\}_{1\le i\le I}$, $G(\theta, \alpha)$ is defined as an *auxiliary function* of $D(\theta)$ if it satisfies

$$D(\theta) = \min_{\alpha} G(\theta, \alpha), \tag{2.6}$$

where we refer to α as *auxiliary variables*.

Theorem 2.1 *$D(\theta)$ is non-increasing under the updates:*

$$\alpha \leftarrow \operatorname*{argmin}_{\alpha} G(\theta, \alpha), \tag{2.7}$$

$$\theta_i \leftarrow \operatorname*{argmin}_{\theta_i} G(\theta, \alpha) \ \ (i = 1, \ldots, I). \tag{2.8}$$

Proof Let us set θ at an arbitrary value $\theta^{(\ell)}$ and define

$$\alpha^{(\ell+1)} = \operatorname*{argmin}_{\alpha} G(\theta^{(\ell)}, \alpha), \quad \theta^{(\ell+1)} = \left\{\operatorname*{argmin}_{\theta_i} G(\theta, \alpha^{(\ell+1)})\right\}_{1 \le i \le I}. \tag{2.9}$$

First, it is obvious that $D(\theta^{(\ell)}) = G(\theta^{(\ell)}, \alpha^{(\ell+1)})$. Next, we can confirm that $G(\theta^{(\ell)}, \alpha^{(\ell+1)}) \ge G(\theta^{(\ell+1)}, \alpha^{(\ell+1)})$. By definition, it is clear that $G(\theta^{(\ell+1)}, \alpha^{(\ell+1)}) \ge D(\theta^{(\ell+1)})$ and so we can finally show that $D(\theta^{(\ell)}) \ge D(\theta^{(\ell+1)})$. A sketch of this proof can be found in Fig. 2.4.

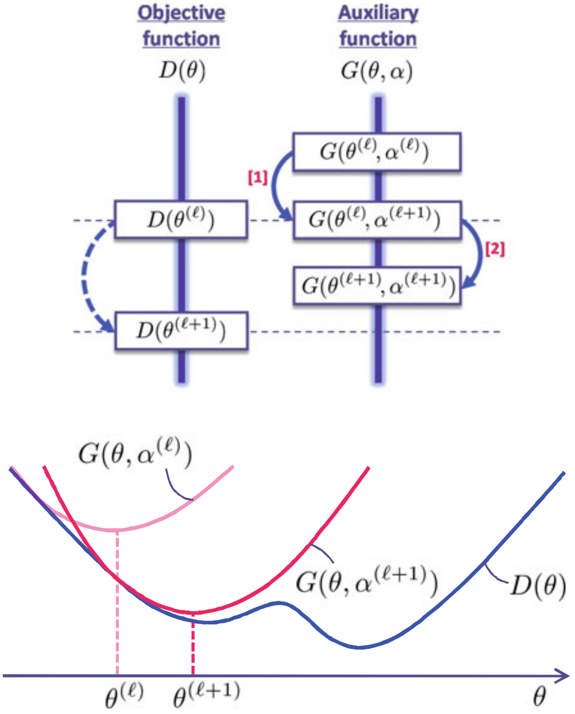
Objective function
$D(\theta)$
Auxiliary function
$G(\theta,\alpha)$
$G(\theta^{(\ell)},\alpha^{(\ell)})$
[1]
$D(\theta^{(\ell)})$
$G(\theta^{(\ell)},\alpha^{(\ell+1)})$
[2]
$G(\theta^{(\ell+1)},\alpha^{(\ell+1)})$
$D(\theta^{(\ell+1)})$
$G(\theta,\alpha^{(\ell)})$
$G(\theta,\alpha^{(\ell+1)})$
$D(\theta)$
$\theta^{(\ell)}$
$\theta^{(\ell+1)}$
θ

Fig. 2.4 Sketch of process of auxiliary function method

2.4.4 NMF Algorithm with Euclidean Distance

By employing the principle of the auxiliary function approach, we first derive an NMF algorithm using $D_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U})$ as the goodness-of-fit criterion. By using $\overset{z}{=}$ to denote equality up to a term independent of z, we can write $D_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U})$ as

$$D_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U}) \overset{H,U}{=} \sum_{k,n} (-2y_{k,n}x_{k,n} + x_{k,n}^2), \tag{2.10}$$

where

$$x_{k,n} = \sum_{m} h_{k,m} u_{m,n}. \tag{2.11}$$

We want to design an auxiliary function such that the matrix elements are separated into individual terms. Note that $x_{k,n}^2$ is a term involving $h_{k,1}, \ldots, h_{k,M}$ and $u_{1,n}, \ldots, u_{M,n}$. Since a quadratic function is convex, we can employ Jensen's inequality to construct a desired auxiliary function.

Theorem 2.2 (Jensen's inequality for convex functions with non-negative arguments (Fig. 2.5)) *For an arbitrary convex function g with I non-negative arguments $z_1, \ldots, z_I$, we have*

$$g\Big(\sum_i z_i\Big) \le \sum_i \lambda_i g\left(\frac{z_i}{\lambda_i}\right), \tag{2.12}$$

where $\lambda_1, \ldots, \lambda_1$ are non-negative weights satisfying $\sum_i \lambda_i = 1$. Equality in this inequality holds when

$$\lambda_i = \frac{z_i}{\sum_j z_j}. \tag{2.13}$$

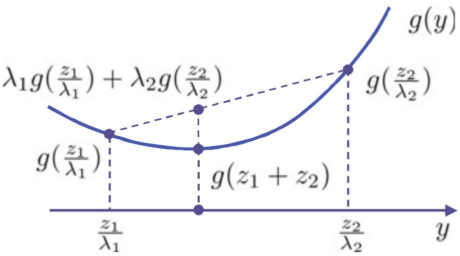
$g(y)$
$\lambda_1 g(\frac{z_1}{\lambda_1}) + \lambda_2 g(\frac{z_2}{\lambda_2})$
$g(\frac{z_2}{\lambda_2})$
$g(\frac{z_1}{\lambda_1})$
$g(z_1 + z_2)$
$\frac{z_1}{\lambda_1}$
$\frac{z_2}{\lambda_2}$
y

Fig. 2.5 Jensen's inequality for functions with non-negative arguments for $I = 2$ case

Since $h_{k,m}u_{m,n} \geq 0$, we can apply this to $x_{k,n}^2$

$$x_{k,n}^2 \leq \sum_m \lambda_{k,m,n} \left(\frac{h_{k,m}u_{m,n}}{\lambda_{k,m,n}} \right)^2, \tag{2.14}$$

where $\lambda_{k,m,n} \geq 0$, $\sum_m \lambda_{k,m,n} = 1$. Here, we notice that the right-hand side of this inequality is given as the sum of terms each relating to $h_{k,m}$ and $u_{m,n}$. It is also important to note that the equality holds when $\frac{h_{k,1}u_{1,n}}{\lambda_{k,1,n}} = \cdots = \frac{h_{k,M}u_{M,n}}{\lambda_{k,M,n}}$, namely

$$\lambda_{k,m,n} = \frac{h_{k,m}u_{m,n}}{x_{k,n}}. \tag{2.15}$$

Hence, the function obtained by replacing the term $x_{k,n}^2$ in $D_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U})$ with the right-hand side of Eq. (2.14)

$$G_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda}) = \sum_{k,n} \left(y_{k,n}^2 - 2y_{k,n} \sum_m h_{k,m}u_{m,n} + \sum_m \frac{h_{k,m}^2 u_{m,n}^2}{\lambda_{k,m,n}} \right) \tag{2.16}$$

satisfies the requirement of an auxiliary function for $D_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U})$. Here, $\boldsymbol{\lambda} = \{\lambda_{k,m,n}\}_{K \times M \times N}$. By using $G_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda})$, we can develop an iterative algorithm for locally minimizing $D_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U})$, that consists of performing

$$\boldsymbol{\lambda} \leftarrow \underset{\boldsymbol{\lambda}}{\operatorname{argmin}}\, G_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda}), \tag{2.17}$$

$$\boldsymbol{H} \leftarrow \underset{\boldsymbol{H}}{\operatorname{argmin}}\, G_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda}), \quad \boldsymbol{U} \leftarrow \underset{\boldsymbol{U}}{\operatorname{argmin}}\, G_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda}). \tag{2.18}$$

First, Eq. (2.17) is given as Eq. (2.15) as mentioned above. Next, Eq. (2.18) must be solved subject to non-negativity. $G_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda})$ is a quadratic function of each matrix element $h_{k,m}$, which can be minimized when

$$\hat{h}_{k,m} = \frac{\displaystyle\sum_n y_{k,n}u_{m,n}}{\displaystyle\sum_n u_{m,n}^2 / \lambda_{k,m,n}}. \tag{2.19}$$

In the same way, $G_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda})$ can be minimized with respect to $u_{m,n}$ when

$$\hat{u}_{m,n} = \frac{\displaystyle\sum_k y_{k,n}h_{k,m}}{\displaystyle\sum_k h_{k,m}^2 / \lambda_{k,m,n}}. \tag{2.20}$$

If these values become negative, the minimizers of $G_{\mathrm{EU}}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda})$ within the non-negativity constraint will obviously be $h_{k,m} = 0$ and $u_{m,n} = 0$. Thus, Eq. (2.18) is given as $h_{k,m} = \max\{\hat{h}_{k,m}, 0\}$ and $u_{m,n} = \max\{\hat{u}_{m,n}, 0\}$. Note, however, that when all the elements of $\boldsymbol{H}$, $\boldsymbol{U}$ and $\boldsymbol{\lambda}$ are non-negative, both (2.19) and (2.20) necessarily become non-negative. Hence, if the initial values of $\boldsymbol{H}$ and $\boldsymbol{U}$ are set at non-negative values, $h_{k,m}$ and $u_{m,n}$ will always be updated to non-negative values. In such a situation, the update equations for $h_{k,m}$ and $u_{m,n}$ can be written simply as $h_{k,m} = \hat{h}_{k,m}$ and $u_{m,n} = \hat{u}_{m,n}$. By substituting Eq. (2.17) into Eq. (2.18), we obtain the following algorithm:

NMF algorithm with the Euclidean distance

1. Set $\boldsymbol{H}$ and $\boldsymbol{U}$ at non-negative values.
2. Repeat the following updates until convergence.

$$h_{k,m} \leftarrow h_{k,m} \frac{\sum_n y_{k,n} u_{m,n}}{\sum_n x_{k,n} u_{m,n}} \qquad u_{m,n} \leftarrow u_{m,n} \frac{\sum_k y_{k,n} h_{k,m}}{\sum_k x_{k,n} h_{k,m}}$$

Since each variable is updated by multiplying the value at the previous iteration by a non-negative factor, this kind of algorithm is often referred to as a “multiplicative update algorithm” [1].

2.4.5 NMF Algorithm with I Divergence

An NMF algorithm using $D_{\mathrm{KL}}(\boldsymbol{H}, \boldsymbol{U})$ as a goodness-of-fit criterion can be derived in a similar way. $D_{\mathrm{KL}}(\boldsymbol{H}, \boldsymbol{U})$ is equal up to a constant term to

$$D_{\mathrm{KL}}(\boldsymbol{H}, \boldsymbol{U}) \overset{H,U}{=} \sum_{k,n} (-y_{k,n} \log x_{k,n} + x_{k,n}). \tag{2.21}$$

Here, $-y_{k,n} \log x_{k,n}$ is a nonlinear term involving $h_{k,1}, \ldots, h_{k,M}$ and $u_{1,n}, \ldots, u_{M,n}$. By using the fact that a negative logarithmic function is convex and $h_{k,m} u_{m,n} \geq 0$, we can apply Theorem 2.2

$$-\log x_{k,n} \leq -\sum_m \lambda_{k,m,n} \log \left(\frac{h_{k,m} u_{m,n}}{\lambda_{k,m,n}} \right)$$

to construct a desired auxiliary function, from which we obtain the following algorithm:

NMF algorithm with generalized KL divergence (*I* divergence)

1. Set $\boldsymbol{H}$ and $\boldsymbol{U}$ at non-negative values.
2. Repeat the following updates until convergence.

$$h_{k,m} \leftarrow h_{k,m} \frac{\sum_n y_{k,n} u_{m,n} \big/ x_{k,n}}{\sum_n u_{m,n}} \qquad u_{m,n} \leftarrow u_{m,n} \frac{\sum_k y_{k,n} h_{k,m} \big/ x_{k,n}}{\sum_k h_{k,m}}$$

2.4.6 NMF Algorithm with IS Divergence

Here, we show an NMF algorithm using the IS divergence as a goodness-of-fit criterion developed by the author in 2006 [9]. By omitting the terms that do not depend on $\boldsymbol{H}$ and $\boldsymbol{U}$, $D_{\mathrm{IS}}(\boldsymbol{H}, \boldsymbol{U})$ is written as

$$D_{\mathrm{IS}}(\boldsymbol{H}, \boldsymbol{U}) \overset{H,U}{=} \sum_{k,n} \left(\frac{y_{k,n}}{x_{k,n}} + \log x_{k,n} \right). \tag{2.22}$$

In a way similar to that described in the previous subsection, we want to design an auxiliary function such that the matrix elements are separated into individual terms. First, by using the fact that the reciprocal function is convex on a positive half-axis, $h_{k,m} u_{m,n} \geq 0$ and $y_{k,n} \geq 0$, we can apply Theorem 2.2 to the term $1/x_{k,n}$

$$\frac{1}{x_{k,n}} \leq \sum_m \lambda_{k,m,n} \left(1 \Big/ \frac{h_{k,m} u_{m,n}}{\lambda_{k,m,n}} \right), \tag{2.23}$$

where $\lambda_{k,m,n}$ is a positive weight satisfying $\lambda_{k,m,n} > 0$ and $\sum_m \lambda_{k,m,n} = 1$. Next, let us focus on the term $\log x_{k,n}$. Since the positive logarithmic function is concave (not convex), the strategy using Jensen's inequality cannot be used. However, we can apply a different inequality as described below. Given a differentiable concave function g, we can show that a tangent line to g at an arbitrary tangent point $\alpha \in \mathbb{R}$ lies entirely above the graph of g, namely for all $x \in \mathbb{R}$,

$$g(x) \leq g(\alpha) + (x - \alpha) g'(\alpha). \tag{2.24}$$

Obviously, the equality of this inequality holds if and only if $\alpha = x$. By applying this to $\log x_{k,n}$, we obtain

$$\log x_{k,n} \leq \log \alpha_{k,n} + \frac{1}{\alpha_{k,n}}(x_{k,n} - \alpha_{k,n}), \tag{2.25}$$

where $\alpha_{k,n}$ is an arbitrary real number. It is important to note that the right-hand side of this inequality is given as a first order function of the matrix elements. Hence, the function obtained by replacing the terms $1/x_{k,n}$ and $\log x_{k,n}$ in $D_{\mathrm{IS}}(\boldsymbol{H}, \boldsymbol{U})$ with the right-hand sides of Eqs. (2.23) and (2.25), such that

$$G_{\mathrm{IS}}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \sum_{k,n} \left(\sum_{m} \frac{y_{k,n}\lambda_{k,m,n}^2}{h_{k,m}u_{m,n}} + \sum_{m} \frac{h_{k,m}u_{m,n}}{\alpha_{k,n}} - \log y_{k,n} + \log \alpha_{k,n} - 2 \right), \tag{2.26}$$

satisfies the requirement of an auxiliary function for $D_{\mathrm{IS}}(\boldsymbol{H}, \boldsymbol{U})$ [9]. Note that the equalities of Eqs. (2.23) and (2.25) hold if and only if

$$\lambda_{k,m,n} = \frac{h_{k,m}u_{m,n}}{x_{k,n}}, \quad \alpha_{k,n} = x_{k,n}. \tag{2.27}$$

By applying Theorem 2.1 and deriving each update equation, we obtain the following algorithm:

NMF algorithm with the IS divergence

1. Set $\boldsymbol{H}$ and $\boldsymbol{U}$ at non-negative values.
2. Repeat the following updates until convergence.

$$h_{k,m} \leftarrow h_{k,m} \left(\frac{\sum_n y_{k,n}u_{m,n}/x_{k,n}^2}{\sum_n u_{m,n}/x_{k,n}} \right)^{1/2} \qquad u_{m,n} \leftarrow u_{m,n} \left(\frac{\sum_k y_{k,n}h_{k,m}/x_{k,n}^2}{\sum_k h_{k,m}/x_{k,n}} \right)^{1/2}$$

2.4.7 NMF Algorithm with β Divergence

The three divergence measures given in Eqs. (2.3)–(2.5) can be described in a unified manner using a criterion called the β divergence [10]

$$\mathcal{D}_\beta(y|x) = y\frac{y^{\beta-1} - x^{\beta-1}}{\beta - 1} - \frac{y^\beta - x^\beta}{\beta}, \tag{2.28}$$

where β is a real number such that $\beta \neq 0$ and $\beta \neq 1$. By using the fact that $\lim_{\beta\to 0}(x^\beta - y^\beta)/\beta = \log(x/y)$, it can be confirmed that Eq. (2.28) reduces to the IS divergence when $\beta \to 0$, the I divergence when $\beta \to 1$ and the Euclidean distance when $\beta = 2$, respectively. Here, we show a generalized NMF algorithm using the β divergence as a goodness-of-fit criterion, that we have previously developed [11]. The first term $(y^{\beta-1} - x^{\beta-1})/(\beta - 1)$ of Eq. (2.28) is convex in x when $\beta \leq 2$ and is concave otherwise. On the other hand, the second term $-(y^\beta - x^\beta)/\beta$ is concave in x when $\beta \leq 1$ and is convex otherwise. In a way similar to the idea of [9], we can construct an auxiliary function by applying Eq. (2.12) to the convex term and Eq. (2.24) to the concave term. By using this auxiliary function, we can derive update equations given in closed form in the same way as in the previous subsections. The NMF algorithm derived using this idea is summarized as follows:

NMF algorithm with the β divergence

1. Set $\boldsymbol{H}$ and $\boldsymbol{U}$ at non-negative values, choose β and set $\varphi(\beta)$ at

$$\varphi(\beta) = \begin{cases} 1/(2-\beta) & (\beta < 1) \\ 1 & (1 \leq \beta \leq 2) \\ 1/(\beta - 1) & (\beta > 2) \end{cases}.$$

2. Repeat the following updates until convergence.

$$h_{k,m} \leftarrow h_{k,m}\left(\frac{\sum_n y_{k,n} x_{k,n}^{\beta-2} u_{m,n}}{\sum_n x_{k,n}^{\beta-1} u_{m,n}}\right)^{\varphi(\beta)} \qquad u_{m,n} \leftarrow u_{m,n}\left(\frac{\sum_k y_{k,n} x_{k,n}^{\beta-2} h_{k,m}}{\sum_k x_{k,n}^{\beta-1} h_{k,m}}\right)^{\varphi(\beta)}$$

It can be readily verified that the above algorithm reduces to the multiplicative update algorithms with the IS divergence, the I divergence and the Euclidean distance presented in Sects. 2.4.4, 2.4.5 and 2.4.6 when $\beta = 0, 1, 2$, respectively.

2.5 Interpretation of NMF as Generative Model

2.5.1 β Divergence Versus Tweedie Distribution

The optimization problems of NMF with the Euclidean distance, I divergence, IS divergence and β divergence are equivalent to the problems of the maximum likelihood estimation of $\boldsymbol{H}$ and $\boldsymbol{U}$, where each element y_{kn} of $\boldsymbol{Y}$ is assumed to have been generated independently from the normal distribution, Poisson distribution, exponential distribution and Tweedie distribution with the mean $x_{k,n}$

$$y_{k,n} \sim \mathcal{N}(y_{k,n}; x_{k,n}, \sigma^2), \tag{2.29}$$
$$y_{k,n} \sim \text{Poisson}(y_{k,n}; x_{k,n}), \tag{2.30}$$
$$y_{k,n} \sim \text{Exponential}(y_{k,n}; x_{k,n}), \tag{2.31}$$
$$y_{k,n} \sim \text{Tweedie}(y_{k,n}; x_{k,n}, \phi), \tag{2.32}$$

respectively, where

$$\mathcal{N}(z; \mu, \sigma^2) = \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-(z-\mu)^2/2\sigma^2}, \tag{2.33}$$
$$\text{Poisson}(z; \mu) = \mu^z e^{-\mu}/z! \;\; (z \geq 0), \tag{2.34}$$
$$\text{Exponential}(z; \mu) = \tfrac{1}{\mu} e^{-z/\mu} \;\; (z \geq 0), \tag{2.35}$$
$$\text{Tweedie}(z; \mu, \phi) = a(z, \phi) e^{\frac{1}{\phi}(z\rho(\mu)-\kappa(\mu))}, \tag{2.36}$$
$$\rho(\mu) = \begin{cases} \frac{\mu^{\beta-1}-1}{\beta-1} & (\beta \neq 1) \\ \log \mu & (\beta = 1) \end{cases}, \;\; \kappa(\mu) = \begin{cases} \frac{\mu^{\beta}-1}{\beta} & (\beta \neq 0) \\ \log \mu & (\beta = 0). \end{cases}$$

This can be confirmed as follows. All the log-likelihoods $L(x_{k,n}) = \log p(y_{k,n}|x_{k,n})$ defined by Eqs. (2.29)–(2.32) are maximized when $x_{k,n} = y_{k,n}$. Thus, $L(y_{k,n}) \geq L(x_{k,n})$. Hence, the log-likelihood differences $L(y_{k,n}) - L(x_{k,n})$ can be regarded as non-negative measures of the dissimilarity between $y_{k,n}$ and $x_{k,n}$ that become 0 only when $x_{k,n} = y_{k,n}$. We can see that the log-likelihood differences $L(y_{k,n}) - L(x_{k,n})$ for Eqs. (2.29)–(2.32) are equal to Eqs. (2.3)–(2.5) and (2.28), respectively.

2.5.2 Bregman Divergence Versus Natural Exponential Family

As we have seen in the four examples above, an assumption regarding the divergence measure for a certain model-fitting problem is associated with a probability density function assumption regarding the observed data. In this subsection, I show that the class of probabilistic distributions belonging to the natural exponential family is associated with the class of goodness-of-fit criteria called the Bregman divergence and that the β divergence is a special case of the Bregman divergence. In the following, I will omit the subscripts k, n for simplicity and assume that an element y of the observed matrix follows a probability distribution belonging to the exponential family

$$y \sim \exp\big\{\eta T(y) - \psi(\eta) + c(y)\big\}, \tag{2.37}$$

where ψ is an infinitely differentiable, strictly convex function. η is called a natural parameter and is a function of the parameters characterizing the distribution. Here, we consider the case $T(y) = y$, whose distribution class is called the natural exponential family.

First, we introduce the Legendre transform of ψ

$$\phi(z) = \max_v (vz - \psi(v)). \tag{2.38}$$

Since ψ is a convex function, ϕ also becomes a convex function due to the property of the Legendre transform. By using v^* to denote $\phi(z)$, i.e., v that maximizes $vz - \psi(v)$, v^* satisfies $(vz - \psi(v))' = 0$, namely

$$\psi'(v^*) = z. \tag{2.39}$$

Next, by using the fact that the cumulant generating function of $y \sim \exp\big\{\eta y - \psi(\eta) + c(y)\big\}$ is given as $K(t) = \log \mathbb{E}[e^{yt}] = \psi(t + \eta) - \psi(\eta)$, we can write $x := \mathbb{E}[y] = K'(0)$ as

$$x = \psi'(\eta). \tag{2.40}$$

Since ψ is a convex function, ψ' is a one-to-one function. Thus, there is a one-to-one correspondence between η and x. By comparing Eq. (2.40) with Eq. (2.39), we can show that $\eta = \underset{v}{\mathrm{argmax}}(vx - \psi(v))$. $\phi(x)$ can thus be written as

$$\phi(x) = \eta(x)x - \psi(\eta(x)). \tag{2.41}$$

Note that here η is written as $\eta(x)$ to emphasize that it is a function of x. Here, by differentiating both sides of Eq. (2.41) with respect to x, we have

$$\phi'(x) = \eta(x) + \eta'(x)x - \psi'(\eta(x))\eta'(x). \tag{2.42}$$

By plugging Eq. (2.40) into Eq. (2.42), the second and third terms cancel each other out, thus resulting in $\phi'(x) = \eta(x)$.

By substituting the two relationships $\phi(x) = \eta x - \psi(\eta)$ and $\phi'(x) = \eta(x)$ given above into the probability density function of the natural exponential family $\exp\{\eta y - \psi(\eta) + c(y)\} = \exp\{\eta x - \psi(\eta) + \eta(y - x) + c(y)\}$, we obtain

$$p(y|x) = \exp\{\phi(x) + \phi'(x)(y - x) + c(y)\}. \tag{2.43}$$

Here, it is important to note that the log-likelihood of x, $L(x) = \log p(y|x) = \phi(x) + \phi'(x)(y - x) + c(y)$, is maximized when $x = y$, since $(\phi(x) + \phi'(x)(y - x))' = \phi''(x)(y - x)$. Thus, $L(y) \geq L(x)$. Hence, the log-likelihood difference $L(y) - L(x)$

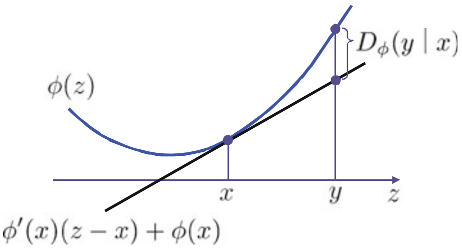

Fig. 2.6 Bregman divergence

$$\mathcal{D}_\phi(y|x) = \phi(y) - \phi(x) - \phi'(x)(y-x), \tag{2.44}$$

can be regarded as a non-negative measure of the dissimilarity between x and y that becomes 0 only when $x = y$. This measure is called the Bregman divergence [12]. As shown in Fig. 2.6, $D_\phi(y|x)$ corresponds to the difference between the convex function ϕ and its tangent line at point x. We can see from this figure that $D_\phi(y|x)$ is always non-negative and that $D_\phi(y|x)$ becomes 0 only when x and y are equal.

The β divergence introduced in Sect. 2.4.7 is a special case of the Bregman divergence with

$$\phi(x) = \begin{cases} -\log x + x - 1 & (\beta = 0) \\ x \log x - x + 1 & (\beta = 1) \\ \frac{x^\beta}{\beta(\beta-1)} - \frac{x}{\beta-1} + \frac{1}{\beta} & (\text{otherwise}) \end{cases} \quad [13]. \tag{2.45}$$

Thus, the Euclidean distance, I divergence, and IS divergence, which are special cases of the β divergence, are also special cases of the Bregman divergence.

An attempt was made by Dhillon and Sra to derive a multiplicative update algorithm for NMF with the Bregman divergence under a limited class of ϕ [14]. However, its generalization to an arbitrary ϕ has yet to be proposed.

2.6 Relation to Probabilistic Latent Semantic Analysis (pLSA)

The concept of probabilistic Latent Semantic Analysis (pLSA) [15], which is a technique that was originally developed for document clustering and indexing, is closely related to NMF. This section describes the relationship between these two techniques.

Let $y_{k,n}$ be the number of times word k occurs in document n. The histogram of all possible words $\boldsymbol{y}_n = (y_{1,n}, \ldots, y_{K,n})^\mathsf{T}$ in document n is called a document data. The number of times a particular word occurs may depend heavily on the topic of the document such as politics, economy, sports, entertainment, and culture. The aim of pLSA is to estimate topics from document data based on this dependence.

Let $p(k|m)$ be the probability that word k occurs when the topic is m and $p(m|n)$ be the probability that the topic of document n is m. Then, the probability $p(k|n)$ that word k occurs in document n can be written as

$$p(k|n) = \sum_{m} p(k|m)p(m|n). \tag{2.46}$$

By putting $x_{k,n} = p(k|n)$, $h_{k,m} = p(k|m)$ and $u_{m,n} = p(m|n)$, and by arranging them in matrices $\boldsymbol{X} = (x_{k,n})_{K\times N}$, $\boldsymbol{H} = (h_{k,m})_{K\times M}$ and $\boldsymbol{U} = (u_{m,n})_{M\times N}$, Eq. (2.46) can be written in matrix notation as $\boldsymbol{X} = \boldsymbol{H}\boldsymbol{U}$. If each word in a set of document data is assumed to be generated independently according to the distribution $p(k|n)$, the probability that all the document data are generated becomes $\prod_{k,n} p(k|n)^{y_{k,n}}$. Since both $H_{k,m} = p(k|m)$ and $U_{m,n} = p(m|n)$ are unknown, the maximum likelihood estimation of $\boldsymbol{H}$ and $\boldsymbol{U}$ can be formulated as an optimization problem of maximizing

$$\log p(\boldsymbol{Y}|\boldsymbol{H}, \boldsymbol{U}) = \sum_{k,n} y_{k,n} \log x_{k,n}, \tag{2.47}$$

with respect to $\boldsymbol{H}$ and $\boldsymbol{U}$ subject to non-negativity and sum-to-one constraints: $h_{k,m} \geq 0$, $\sum_k h_{k,m} = 1$, $u_{m,n} \geq 0$, $\sum_m u_{m,n} = 1$. By comparing Eqs. (2.47) and (2.21), we notice that the above log-likelihood is exactly opposite to the first term of Eq. (2.21). Furthermore, as the second term of Eq. (2.21) can be seen as corresponding to a Lagrange multiplier term for $x_{k,n}$, the pLSA optimization problem has the same form as that of NMF with the I divergence criterion. Indeed, it turns out that the optimization algorithm described in Sect. 2.4.5 is equivalent to the expectation-maximization (EM) algorithm obtained by treating the topic index m as a latent variable up to the normalization of $\boldsymbol{H}$ and $\boldsymbol{U}$.

As described above, the way in which the likelihood function of pLSA is defined is different from NMF described in Sect. 2.5. While pLSA treats $h_{k,m}$ and $u_{m,n}$ as probability distributions over k and m, NMF treats them as random variables. Namely, pLSA is categorized as mixture models (models defined as the sum of probability distributions) whereas NMF is categorized as factor models (models defined as the distribution of the sum of random variables). The Bayesian extension of pLSA is called the latent Dirichlet allocation (LDA) [16] and the Bayesian extension of NMF with the I divergence criterion is discussed for example in [17].

2.7 Applications to Audio Signal Processing Problems

2.7.1 Audio Source Separation and Music Transcription

Smaragdis and Brown proposed an automatic music transcription method that uses NMF to decompose the magnitude (or power) spectrograms of music signals into spectrograms associated with individual pitches [18]. With this approach, the magnitude (or power) spectrogram of a mixture signal, interpreted as a non-negative matrix $\boldsymbol{Y}$, is factorized into the product of two non-negative matrices $\boldsymbol{H}$ and $\boldsymbol{U}$ (See Fig. 2.1). This can in turn be interpreted as approximating the observed spectra at each time frame as a linear sum of basis spectra scaled by time-varying amplitudes, and amounts to decomposing the observed spectrogram into the sum of rank-1 spectrograms. As described in Sect. 2.3, an important feature of NMF is that its non-negativity constraint usually induces sparse representations, i.e., $\boldsymbol{U}$ with a relatively large number of zero entries. This means that each observed spectrum is parsimoniously represented using only a few active basis spectra. In such situations, the sequence of observed spectra can be approximated reasonably well when each basis spectrum expresses the spectrum of an underlying audio event that occurs frequently over the entire observed range. Thus, with music signals, each basis spectrum usually becomes the spectrum of a frequently used pitch in the music piece under analysis.

This approach is based on two assumptions; one is that magnitude (or power) spectra are additive and the other is that the magnitude spectrum of each sound source is constant up to the scale over time (i.e., only the scale of the spectrum is time-variant). However, these assumptions do not hold in reality. This section introduces some of the attempts that have been made to develop variants of NMF that aim to relax these assumptions.

2.7.2 Complex NMF

Audio signals in the time domain (sound waves) are additive. Since typical methods for time-frequency decomposition, such as the short-time Fourier transform (STFT) and the wavelet transform, are linear, complex spectrograms of audio signals are also additive. However, since the transformation of complex spectrograms into magnitude (or power) spectrograms is nonlinear, magnitude spectrograms are non-additive. Namely, the magnitude spectrum of the sum of two waveforms is not equal to the sum of the magnitude spectra of the two waveforms. This implies that decomposing a magnitude spectrogram into the sum of additive components does not necessarily lead to an appropriate decomposition of the audio signal.

To address this shortcoming of the NMF approach, I previously proposed a framework called the “Complex NMF” [19], which makes it possible to realize NMF-like signal decompositions in the complex spectrogram domain. The key idea behind the NMF approach was to model the magnitude spectrogram of a mixture signal as the sum of rank-1 magnitude spectrograms. By contrast, the key idea behind the proposed approach is to model the complex spectrogram of a mixture signal as the sum of complex spectrograms each having a rank-1 structure in the magnitude domain. This idea can be formulated as follows.

Let $a_{m,k,n} \in \mathbb{C}$ denote the complex spectrogram of source m. The complex spectrogram of a mixture signal consisting of M sources is given as

$$f_{k,n} = \sum_{m=1}^{M} a_{m,k,n} = \sum_{m} |a_{m,k,n}| e^{j\phi_{m,k,n}}, \tag{2.48}$$

where $\phi_{m,k,n}$ denotes the phase spectrogram of source m. Here, if we assume that the magnitude spectrogram of each source has a rank-1 structure, we can write $|a_{m,k,n}| = h_{k,m} u_{m,n}$. This leads to a complex spectrogram model of the form:

$$f_{k,n} = \sum_{m} h_{k,m} u_{m,n} e^{j\phi_{m,k,n}}. \tag{2.49}$$

It is important to emphasize that $\phi_{m,k,n}$ is indexed by n, meaning that this model allows the phase spectrum of each source to vary freely over time. The aim of Complex NMF is to fit this model to an observed complex spectrogram through the estimation of H, U and ϕ. It should be noted that unlike NMF, this model allows the components to cancel each other out (since the real and imaginary parts of the complex spectrum of each source can take either positive or negative values), and so when there are no constraints, it does not naturally produce sparse representations. Thus, to obtain sparse representations similar to NMF, some constraint is needed to induce the sparsity of $\boldsymbol{U}$. In [19], I formulated an optimization problem of minimizing

$$I(H, U, \phi) := \sum_{k,n} \left| y_{k,n} - f_{k,n} \right|^2 + 2\gamma \sum_{m,n} |u_{m,n}|^p, \tag{2.50}$$

with respect to H, U and ϕ where the second term is a sparse regularization term, and derived an iterative algorithm based on the auxiliary function approach. Here, $0 < p < 2$ and $\gamma \geq 0$ are constants. The main difficulty with this optimization problem lies in the nonlinear interdependence of $\phi_{1,k,n}, \ldots, \phi_{M,k,n}$ and the discontinuity of the gradients with respect to $u_{m,n}$. The nonlinear interdependence of $\phi_{1,k,n}, \ldots, \phi_{M,k,n}$ arises from the "square-of-sum" form in the first term of Eq. (2.50). To derive closed-form update equations using the auxiliary function approach in a similar way to Sect. 2.4.4, it is desirable to design an upper bound function that has a "sum-of-squares" form for this term. However, unlike Sect. 2.4.4, Theorem 2.2 cannot be applied in this case, since $h_{k,m} u_{m,n} e^{j\phi_{m,k,n}}$ is a complex number. Instead, in [19] I proposed invoking the following inequality:

Theorem 2.3 (Jensen's inequality for convex functions with complex arguments) *For an arbitrary convex function g with complex arguments y and $z_1, \ldots, z_I$, we have*

$$g\Big(y - \sum_{i} z_i\Big) \leq \sum_{i} \beta_i g\left(\frac{\alpha_i - z_i}{\beta_i}\right), \tag{2.51}$$

where $\alpha_1, \ldots, \alpha_1$ *are complex variables satisfying* $\sum_i \alpha = y$ *and* $\beta_1, \ldots, \beta_1$ *are positive weights satisfying* $\sum_i \beta = 1$*. Equality in this inequality holds when*

$$\alpha_i = z_i + \beta_i \Big(y - \sum_i z_i\Big). \tag{2.52}$$

Proof Since $\sum_i \alpha = y$, we can write $g(y - \sum_i z_i) = g(\sum_i (\alpha_i - z_i))$. By using arbitrary positive weights $\beta_1, \ldots, \beta_1$ that sum to one, we obtain

$$\begin{aligned} g\Big(\sum_i (\alpha_i - z_i)\Big) &= g\left(\sum_i \beta_i \frac{\alpha_i - z_i}{\beta_i}\right) \\ &\leq \sum_i \beta_i g\left(\frac{\alpha_i - z_i}{\beta_i}\right), \end{aligned} \tag{2.53}$$

where the second line follows from Jensen's inequality. Note that equality in this inequality holds when

$$\frac{\alpha_1 - z_1}{\beta_1} = \cdots = \frac{\alpha_I - z_I}{\beta_I}. \tag{2.54}$$

Letting Z denote the value of Eq. (2.54), α_i is given as $\alpha_i = z_i + \beta_i Z$. Since α_i must sum to y, i.e., $\sum_i \alpha_i = \sum_i z_i + Z = y$, Z is given by $Z = y - \sum_i z_i$. By substituting this into $\alpha_i = z_i + \beta_i Z$, we finally obtain Eq. (2.52).

As for the second term of Eq. (2.50), which is non-differentiable with respect to $u_{m,n}$, we can use the fact that, when $0 < p \leq 2$,

$$|u_{m,n}|^p \leq \frac{p|v_{m,n}|^{p-2}}{2} u_{m,n}^2 + \frac{2-p}{2} |v_{m,n}|^p, \tag{2.55}$$

to construct an upper bound function. Altogether, we obtain an auxiliary function

$$\begin{aligned} I^+(H, U, \phi, \alpha, V) := &\sum_{k,n,m} \frac{1}{\beta_{m,k,n}} \Big| \alpha_{m,k,n} - h_{k,m} u_{m,n} e^{\mathrm{j}\phi_{m,k,n}} \Big|^2 \\ &+ \gamma \sum_{m,n} \Big\{ p|v_{m,n}|^{p-2} {u_{m,n}}^2 + (2-p)|v_{m,n}|^p \Big\}, \end{aligned} \tag{2.56}$$

which has a "sum-of-squares" form. Here, $\beta_{m,k,n}$ is a positive weight that can be set arbitrarily subject to $\sum_m \beta_{m,k,n} = 1$. $\alpha_{m,k,n}$ and $v_{m,n}$ are auxiliary variables satisfying $\sum_m \alpha_{m,k,n} = y_{k,n}$. By using this, we can develop a convergence-guaranteed iterative algorithm with closed-form update equations.

2.7.3 Itakura-Saito NMF

Although the additivity of power spectra does not generally hold as mentioned above, it holds in the expectation sense if the signals are assumed to be samples drawn independently from stochastic processes.

If each underlying source signal in a mixture signal within a short-term segment is assumed to have been generated from a zero-mean circularly-stationary Gaussian process, each frequency component of the discrete Fourier transform of that segment independently follows a zero-mean complex normal distribution. Namely, if we let $s_{m,k,n}$ be a component of frequency k of source signal m within segment n (i.e., the complex spectrogram of source m), $s_{m,k,n}$ follows a zero-mean complex normal distribution

$$s_{m,k,n} \sim \mathcal{N}_{\mathbb{C}}\big(s_{m,k,n}; 0, \nu_{m,k,n}\big), \tag{2.57}$$

with variance $\nu_{m,k,n}$, where $\mathcal{N}_{\mathbb{C}}(z; \mu, \nu) = \frac{1}{\pi\nu} e^{-|z-\mu|^2/\nu}$. Note that $\nu_{m,k,n}$ corresponds to the expectation of the power spectrogram of source m, i.e., $\nu_{m,k,n} = \mathbb{E}[|s_{m,k,n}|^2]$. Now, if we assume that the complex spectrogram $y_{k,n}$ of an observed signal is given as $y_{k,n} = \sum_m s_{m,k,n}$, and that $s_{m,k,n}$ and $s_{m',k,n}$ $(m \neq m')$ are statistically independent, $y_{k,n}$ also follows a zero-mean complex normal distribution

$$y_{k,n} \sim \mathcal{N}_{\mathbb{C}}\Big(y_{k,n}; 0, \sum_m \nu_{m,k,n}\Big), \tag{2.58}$$

with variance $\sum_m \nu_{m,k,n}$. By putting $x_{k,n} = \sum_m \nu_{m,k,n}$, the log-likelihood of $x_{k,n}$ given an observation $y_{k,n}$ can be written as

$$L(x_{k,n}) = -\log\pi x_{k,n} - \frac{|y_{k,n}|^2}{x_{k,n}}. \tag{2.59}$$

Since this log-likelihood reaches maximum only when $x_{k,n} = |y_{k,n}|^2$, we have $L(|y_{k,n}|^2) \geq L(x_{k,n})$. We notice that the log-likelihood difference$L(|y_{k,n}|^2) - L(x_{k,n}) \geq 0$ is actually equal to the IS divergence between $|y_{k,n}|^2$ and $x_{k,n}$, i.e., $\mathcal{D}_{\mathrm{IS}}(|y_{k,n}|^2 |x_{k,n})$. Thus, if we assume the expectation of the power spectrogram of each source to have a rank-1 structure, i.e., $\nu_{m,k,n} = h_{k,m}u_{m,n}$, the maximum likelihood estimation of $\boldsymbol{H} = (h_{k,m})_{K\times M}$ and $\boldsymbol{U} = (u_{m,n})_{M\times N}$ is equivalent to the problem of NMF with the IS divergence criterion applied to the observed power spectrogram $|y_{k,n}|^2$ [4, 5].

2.7.4 NMF with Time-Varying Bases

When applying NMF to music spectrograms, we may expect the magnitude spectra of a single musical note produced by an instrument to be represented using a single basis spectrum scaled by time-varying amplitudes. However, its variations in time are actually much richer. For example, a piano note would be more accurately characterized by a succession of several basis spectra corresponding to, for example, "attack," "decay," "sustain" and "release" segments. As another example, singing voices and string instruments have a particular musical effect, vibrato, which can be characterized by its "depth" (the range of pitch variation), and its "speed" (the rate at which the pitch varies). Several variants of NMF have been proposed to represent time-varying spectra by introducing the concept of time-varying bases [20–22].

One approach involves extending NMF to a convolutional version, which finds a decomposition of $\boldsymbol{Y}$ as

$$y_{k,n} \simeq x_{k,n} = \sum_{m}\sum_{l} h_{k,m,l} u_{m,n-l}, \tag{2.60}$$

where $h_{k,m,l}$ can be interpreted as the local time-frequency pattern of the mth audio event and $u_{m,n}$ represents its temporal activation. Since the problem at hand is to decompose the convolutive mixture, this approach is called "non-negative matrix factor deconvolution (NMFD)" [20].

NMFD assumes that the spectrum of each audio event evolves in time in exactly the same way every time it occurs. However, the speeds of the temporal variations are unlikely to be the same all the time. To cope with the varying speeds of the temporal variations of spectra, we proposed modeling the magnitude spectrogram of a mixture signal based on a factorial hidden Markov model (FHMM) formulation [22]. The idea is to model the spectrogram of a mixture signal as the sum of the outputs emitted from multiple HMMs, each representing the spectrogram of an underlying audio event (see Fig. 2.7). Thus, the problem is to find a decomposition of $\boldsymbol{Y}$ as

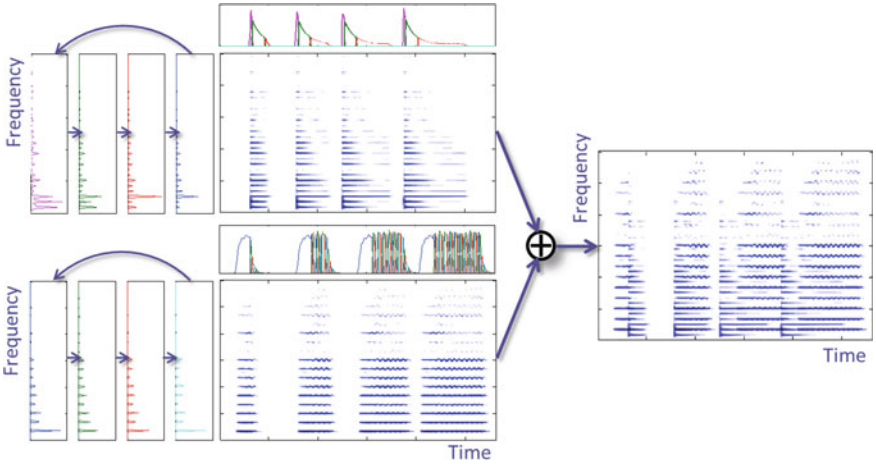

Fig. 2.7 Illustration of the factorial HMM approach [22]. The spectrogram of a mixture signal is modeled as the sum of the outputs emitted from multiple HMMs, each representing the spectrogram of an underlying audio event

$$y_{k,n} \simeq x_{k,n} = \sum_{m} h_{k,m}^{(z_{m,n})} u_{m,n}, \tag{2.61}$$

where $h_{k,m}^{(i)}$ denotes the basis spectrum at state i and $z_{m,n} \in \{1, \dots, I_m\}$ denotes a state variable indicating which basis spectrum is activated at time n. The path of the state variables $z_{m,1}, \dots, z_{m,N}$ is governed by a state transition probability $p(z_{m,n} = a | z_{m,n-1} = b) = \pi_{m,a,b}$.

2.7.5 *Other NMF Variants*

A number of constrained and regularized variants of NMF have been proposed specifically with the aim of solving audio source separation problems. Some examples are as follows. Virtanen proposed incorporating a temporal continuity constraint in the factorization process [23]. Raczyński proposed constraining each basis spectrum so that it had a harmonic structure [24]. Several groups (including mine) independently proposed combining the source-filter model with the NMF model [25–28]. I proposed incorporating a subprocess that involved clustering timbrally similar basis spectra in the factorization process [29].

2.7.6 *Other Applications*

NMF has found several interesting audio-related applications including speech enhancement [30], bandwidth extension [31], singing voice separation [32], drum sound extraction [33], formant tracking [35], echo cancellation [36], and the temporal decomposition of speech [37]. I proposed a blind dereverberation method inspired by the NMF algorithm in [27]. Multichannel extensions of NMF have been proposed independently by several groups (including mine) with an expectation that the modeling concept of NMF can also be useful for multichannel audio source separation problems [39–43].

2.8 Bayesian Nonparametric NMF

2.8.1 *Determination of Basis Number*

The determination of the number of bases is an important issue in NMF. Cemgil and Schmidt proposed formulating the problem of the basis number estimation for NMF as a model selection problem [17, 44]. By using $\boldsymbol{H}^{(M)}$ and $\boldsymbol{U}^{(M)}$ to denote the basis and coefficient matrices with M bases, the marginal likelihood can be thought of as the likelihood function of M, since

$$p(\boldsymbol{Y}|M) = \iint p(\boldsymbol{Y}|\boldsymbol{H}^{(M)}, \boldsymbol{U}^{(M)}) p(\boldsymbol{H}^{(M)}|M) p(\boldsymbol{U}^{(M)}|M) \mathrm{d}\boldsymbol{H}^{(M)} \mathrm{d}\boldsymbol{U}^{(M)}. \quad (2.62)$$

As the exact marginal likelihood involves intractable integrals, some approximation of the log marginal likelihood is usually used as a criterion for model selection. The Bayesian Information Criterion (BIC) [45] and the variational Bayesian lower bound [46] are examples of such approximations. To determine the number of bases with model selection, we need to perform NMF with all possible M settings and find the best model structure by comparing the values of model selection criteria. Although this approach is indeed principled and well founded, such procedures can be time-consuming. By contrast, a framework called the Bayesian nonparameteric approach allows us to avoid performing an explicit model selection procedure and instead reduce this task to a single run of a parameter inference algorithm. In the following, we briefly show some examples of attempts that have been made to apply the Bayesian nonparameteric approach to NMF.

2.8.2 Beta Process NMF and Gamma Process NMF

A Bayesian nonparametric model is a Bayesian model on an infinite-dimensional parameter space. The Bayesian nonparameteric approach refers to a parameter inference framework for Bayesian nonparametric models, which makes it possible to infer the model complexity along with the model parameters by finding a minimal subset of parameters that can explain given observed data.

Bayesian nonparametric models (also known as infinite models) are typically described using stochastic processes. Up to now, many types of infinite models including infinite mixture models and infinite factor models have been proposed in the literature. For instance, infinite counterparts of mixture models, such as the Gaussian mixture model (GMM), hidden Markov model (HMM), probabilistic context-free grammar (PCFG), and probabilistic Latent Semantic Analysis (pLSA), can be constructed using a stochastic process called the Dirichlet process (DP) or its variants. While a Dirichlet distribution is a probabilistic distribution over a finite set of non-negative numbers that sum to 1, the Dirichlet process can be thought of as an extension of it to an infinite set. Thus, the Dirichlet process is a generative model of a categorical distribution (probabilities of discrete random variables) with an infinite dimension, i.e., $\pi_1, \pi_2, \ldots \pi_\infty \in [0, 1]$ satisfying $\sum_{i=1}^{\infty} \pi_i = 1$, which can be used, for example, as a prior distribution over the mixture weights of mixture models. An important property of the Dirichlet process is its sparsity-inducing effect. The categorical distributions generated from a Dirichlet process tend to become sparse. Owing to this property, we can find a minimal subset of mixture components with non-zero weights that explains given observed data through parameter inference.

This is why the use of an infinite mixture model allows us to infer the adequate model complexity (the number of mixture components) from observed data. As mentioned in Sect. 2.6, pLSA can be understood as a particular case of NMF, where the number of mixture components (i.e., the latent topics) corresponds to the basis number. Thus, in a way similar to the NMF approach, it is technically possible to apply pLSA to a magnitude spectrogram by regarding it as document data, where the frequency and time indices are interpreted as the word and document indices, respectively [47], and an infinite counterpart of this approach can be constructed using a Dirichlet process [48]. On the other hand, infinite counterparts of factor models, such as NMF and Independent Component Analysis (ICA), can be constructed using stochastic processes called the beta process (BP) or gamma process (GP). Simply put, the beta process is a generative model of infinitely many variables within the range [0, 1], $\pi_1, \pi_2, \ldots, \pi_\infty \in [0, 1]$, and the gamma process is a generative model of infinitely many non-negative variables, $\theta_1, \theta_2, \ldots, \theta_\infty \in [0, \infty)$. An infinite extension of NMF can be constructed using these stochastic processes. When using the beta process, we introduce a binary variable $z_{m,n} \in \{0, 1\}$ indicating whether the m-th basis exists in data n, with $z_{m,n} = 1$ if data n has a basis m and 0 otherwise. By using $z_{m,n}$, we define $x_{k,n}$ as $x_{k,n} = \sum_{m=1}^{\infty} z_{m,n} h_{k,m} u_{m,n}$ and place a beta process prior $\pi_{m,n} = p(z_{m,n} = 1)$ over $z_{1,n}, \ldots, z_{\infty,n}$ [49, 50]. An important feature of the beta process is its sparsity-inducing effect. The variables generated from a beta process tend to become sparse (most of the variables become almost 0). Owing to this property, we can find a minimal subset of bases that explains given observed data through parameter inference. When using the gamma process, we introduce a non-negative variable $\theta_m \in \mathbb{R}^{\geq 0}$ indicating the contribution made by basis m. By using θ_m, we define $x_{k,n}$ as $x_{k,n} = \sum_{m=1}^{\infty} \theta_m h_{k,m} u_{m,n}$, put some constraint on the scales of $h_{k,m}$ and $u_{m,n}$ (e.g., $\mathbb{E}[h_{k,m}] = 1$ and $\mathbb{E}[u_{m,n}] = 1$), and place a gamma process prior over $\theta_1, \ldots, \theta_\infty$ [22, 51]. An important feature of the gamma process is its sparsity-inducing effect as with the beta process. The variables generated from a gamma process tend to become sparse (most of the variables become almost 0). Owing to this property, we can find a minimal subset of bases that explains given observed data through parameter inference.

2.9 Summary

This chapter described some basic properties of NMF, effects induced by the non-negative constraints, how to derive an iterative algorithm for NMF, some attempts that have been made to apply NMF to audio processing problems, and extensions to the Bayesian nonparametric framework. Readers are referred to other review articles such as [52–55] for further details.

References

1. Lee, D. D., & Seung, H. S. (2000). Algorithms for nonnegative matrix factorization. In *Advances in NIPS* (pp. 556–562).
2. Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*, 111–126.
3. Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, *3*(1), 146–158.
4. Parry, R. M., & Essa, I. (2007). Phase-aware non-negative spectrogram factorization. In Proceedings of ICA (pp. 536–543).
5. Févotte, C., Bertin, N., & Durrieu, J.-L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, *21*(3), 793–830.
6. Ortega, J. M., & Rheinboldt, W. C. (1970). *Iterative solutions of nonlinear equations in several variables*. New York: Academic Press.
7. Hunter, D. R., & Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, *9*, 60–77.
8. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *39*(1), 1–38.
9. Kameoka, H., Goto, M., & Sagayama, S. (2006, August). Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelopes. IPSJ Technical Report (vol. 2006-MUS-66, pp. 77–84) (in Japanese).
10. Eguchi, S., & Kano, Y. (2001). "Robustifying maximum likelihood estimation. Technical Report, Institute of Statistical Mathematics. Research Memo. 802.
11. Nakano, M., Kameoka, H., Le Roux, J., Kitano, Y., Ono, N., & Sagayama, S. (2010). Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence. In *Proceedings of MLSP* (pp. 283–288).
12. Bregman, L. M. (1967). The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*(3), 210–217.
13. Hennequin, R., David, B., & Badeau, R. (2011). Beta-divergence as a subclass of Bregman divergence. *IEEE Signal Processing Letters*, *18*(2), 83–86.
14. Dhillon, I. S., & Sra, S. (2005). Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in NIPS* (pp. 283–290).
15. Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of UAI* (pp. 289–296).
16. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. (J. Lafferty (Ed.)).
17. Cemgil, A. T. (2008). Bayesian inference for nonnegative matrix factorization models, Technical Report CUED/F-INFENG/TR.609, University of Cambridge.
18. Smaragdis, P., & Brown, J. C. (2003). Non-negative matrix factorization for music transcription. In *Proceedings of WASPAA* (pp. 177–180).
19. Kameoka, H., Ono, N., Kashino, K., & Sagayama, S. (2009) Complex NMF: A new sparse representation for acoustic signals. In *Proceedings of ICASSP* (pp. 3437–3440).
20. Smaragdis, P. (2004). Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Proceedings of ICA* (pp. 494–499).
21. Ozerov, A. Févotte, C., & Charbit, M. (2009). Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In *Proceedings of WASPAA* (pp. 121–124).
22. Nakano, M., Le Roux, J., Kameoka, H., Nakamura, T., Ono, N., & Sagayama, S. (2011). Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model. In *Proceedings of WASPAA* (pp. 325–328).

23. Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(3), 1066–1074.
24. Raczynski, S. A., Ono, N., & Sagayama, S. (2007). Multipitch analisys with harmonic non-negative matrix approximation. In *Proceedings of ISMIR* (pp. 381–386).
25. Virtanen, T., & Klapuri, A. (2006). Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In *Advances of NIPS*.
26. Vincent, E., Bertin, N., & Badeau, R. (2008) Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In *Proceedings of ICASSP* (pp. 109–112).
27. Kameoka, H., & Kashino, K. (2009). Composite autoregressive system for sparse source-filter representation of speech. In *Proceedings of ISCAS* (pp. 2477–2480).
28. Yoshii, K., & Goto, M. (2012, October). Infinite composite autoregressive models for music signal analysis. In *Proceedings of The 13th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 79–84).
29. Kameoka, H., Nakano, M., Ochiai, K., Imoto, Y., Kashino, K., & Sagayama, S. (2012). Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints. In *Proceedings of ICASSP* (pp. 5365–5368).
30. Smaragdis, P., Raj, B., & Shashanka, M. V. (2007). Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Proceedings of ICA* (pp. 414–421).
31. Smaragdis, P., & Raj, B. (2007). Example-driven bandwidth expansion. In *Proceedings of WASPAA* (pp. 135–138).
32. Durrieu, J.-L., Richard, G., David, B., & Févotte, C. (2010). Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 564–575.
33. Helén, M., & Virtanen, T. (2005). Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proceedings of EUSIPCO*.
34. Hurmalainen, A., Gemmeke, J., & Virtanen, T. (2011). Non-negative matrix deconvolution in noise robust speech recognition. In *Proceddings of ICASSP* (pp. 4588–4591).
35. Durrieu, J. -L., Thiran, J. -P. (2011). Sparse non-negative decomposition of speech power spectra for formant tracking. In *Proceedings of ICASSP* (pp. 5260–5263).
36. Togami, M., Kawaguchi, Y., Kokubo, H., & Obuchi, Y. (2010). Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization. In *Proceedings of APSIPA* (pp. 522–525).
37. Hiroya, S. (2013). Non-negative temporal decomposition of speech parameters by multiplicative update rules. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(10), 2108–2117.
38. Kameoka, H., Nakatani, T., & Yoshioka, T. (2009). Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms. In *Proceedings of ICASSP* (pp. 45–48).
39. Ozerov, A., & Févotte, C. (2010). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 550–563.
40. Kitano, Y., Kameoka, H., Izumi, Y., Ono, N., & Sagayama, S. (2010). A sparse component model of source sinals and its application to blind source separation. In *Proceedings of ICASSP* (pp. 4122–4125).
41. Sawada, H., Kameoka, H., Araki, S., & Ueda, N. (2011). New formulations and efficient algorithms for multichannel NMF. In *Proceedings of WASPAA* (pp. 153–156).
42. Sawada, H., Kameoka, H., Araki, S., & Ueda, N. (2012). Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization. In *Proceedings of ICASSP* (pp. 261–264).
43. Higuchi, T., Takeda, H., Nakamura, T., Kameoka, H. (2014). A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden Markov models. In *Proceedings of The 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)* (pp. 850–854).

44. Schmidt, M. N., Winther, O., & Hansen, L. K. (2009). Bayesian non-negative matrix factorization. In *Proceedings of ICA* (pp. 540–547).
45. Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.
46. Corduneanu, A., & Bishop, C. M. (2001). Variational Bayesian model selection for mixture distributions. In *Proceedings of AISTATS* (pp. 27–34).
47. Smaragdis, P., Raj, B., & Shashanka, M. (2006). A probabilistic latent variable model for acoustic modeling. In *Advances in NIPS*.
48. Yoshii, K., & Goto, M. (2012). A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(3), 717–730.
49. Knowles, D., & Ghahramani, Z. (2007). Infinite sparse factor analysis and infinite independent components analysis.
50. Liang, D., Hoffman, M. D., & Ellis, D. P. W. (2013). Beta process sparse nonnegative matrix factorization for music.
51. Hoffman, M., Blei, D. & Cook, P. (2010). Bayesian nonparametric matrix factorization for recorded music. In *Proceedings of ICML* (pp. 439–446).
52. Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S. (2009). *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. London: Wiley.
53. Kameoka, H. (2012). Non-negative matrix factorization with application to audio signal processing. *Acoustical Science and Technology*, *68*(11), 559–565. (in Japanese).
54. Sawada, H. (2012). Nonnegative matrix factorization and its applications to data/signal analysis. *IEICE Journal*, *95*, 829–833.
55. Smaragdis, P., Fevotte, C., Mysore, G., Mohammadiha, N., & Hoffman, M. (2014). Static and dynamic source separation using nonnegative factorizations: A unified view. In *IEEE Signal Processing Magazine* (pp. 66–75).