Institut Mines-Télécom

Nonnegative matrix factorization

Roland Badeau
roland.badeau@telecom-paris.fr

Contents

Roland Badeau `roland.badeau@telecom-paris.fr`

List of Figures

Acronyms

AR *Autoregressive*

ARMA *Autoregressive Moving Average*

DNN *Deep Neural Networks*

EM *Expectation-Maximization*

ERB *Equivalent Rectangular Bandwidth*

EUC *Euclidean*

IS *Itakura-Saito*

KL *Kullback-Leibler*

MA *Moving Average*

MAP *Maximum a Posteriori*

ML *Maxmimum Likelihood*

MM *Majorization-Minimization*

MMSE *Minimum Mean Square Error*

NMF *Non-negative Matrix Factorization*

PLCA *Probabilistic Latent Component Analysis*

SAGE *Space Alternating Generalized EM*

STFT *Short Time Fourier Transform*

Mathematical notation

$\mathbb{R}$ set of real numbers

$\mathbb{C}$ set of complex numbers

x (normal font, lower case) scalar

$\boldsymbol{x}$ (bold font, lower case) vector

$\boldsymbol{A}$ (bold font, upper case) matrix

$\|.\|_2$ Euclidean norm of a real vector, or Hermitian norm of a complex vector

$\|.\|_F$ Frobenius norm of a matrix

$.^\top$ transpose of a matrix

$.^H$ conjugate transpose of a matrix

$.^\dagger$ pseudo-inverse of a matrix (if $\boldsymbol{A} \in \mathbb{R}^{M \times K}$ with $M \geq K$, $\boldsymbol{A}^\dagger = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top$)

trace(.) trace of a matrix

diag(.) diagonal matrix formed from a vector of diagonal coefficients, or from a matrix with same diagonal entries

$\boldsymbol{I}_K$ $K \times K$ identity matrix

$\mathbb{E}[.]$ expected value of a random variable or vector

$\mathbb{H}[.]$ entropy of a random variable or vector

$\mathbb{I}[.]$ mutual information of the entries of a random vector

$\hat{h}(\nu) = \sum_{t \in \mathbb{Z}} h(t) e^{-2\imath\pi\nu t}$ discrete time Fourier transform

$L^\infty(\mathbb{R}^M)$ Lebesgue space of essentially bounded functions on $\mathbb{R}^M$

$*$ convolution product between two sequences (scalars, but also matrices and vectors of appropriate dimensions)

Chapter 1

Nonnegative matrix factorization

Non-negative Matrix Factorization (NMF) refers to a set of techniques that have been used to model the spectra of sound sources in various audio applications, including source separation. Sound sources have a structure in time and frequency: music consists of basic units like notes and chords played by different instruments, speech consists of elementary units such as phonemes, syllables or words, and environmental sounds consist of sound events produced by various sound sources. NMF models this structure by representing the spectra of sounds as a sum of components with fixed spectrum and time-varying gain, so that each component in the model represents these elementary units in the sound.

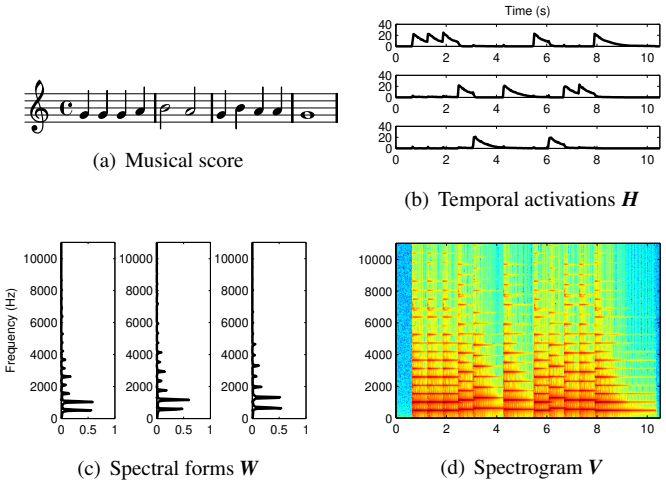
Modeling this structure is beneficial in source separation, since inferring the structure makes it possible to use contextual information for source separation. NMF is typically used to model the magnitude or power spectrogram of audio signals, and its ability to represent the structure of audio sources makes separation possible even in single-channel scenarios.

This chapter presents the use of NMF-based single-channel techniques. In Section 2, several deterministic and probabilistic frameworks for NMF are presented, along with various NMF algorithms. In Section 3, some advanced NMF models are introduced, including regularizations and nonstationary models. Finally, Section 4 summarizes the key concepts introduced in this chapter.

1 Introduction

The NMF was introduced by Lee and Sung to decompose non-negative two-dimensional data into a linear combination of elements in a dictionary [Lee and Seung, 1999].

Given a data matrix $\boldsymbol{V}$ of dimensions $F \times N$ whose coefficients are non-negative, the NMF problem consists in calculating an approximation $\widehat{\boldsymbol{V}}$ of matrix $\boldsymbol{V}$ truncated at rank $K < \min(F, N)$, expressed as a product $\widehat{\boldsymbol{V}} = \boldsymbol{W}\boldsymbol{H}$, where the two matrices $\boldsymbol{W}$ of dimensions $F \times K$ and $\boldsymbol{H}$ of dimensions $K \times N$ have non-negative entries. The columns of the matrix $\boldsymbol{W}$ form the elements of the dictionary and the rows of $\boldsymbol{H}$ contain the coefficients of the decomposition. The dimension K is generally chosen such that $FK + KN \ll FN$, so as to reduce the dimension of the data. The NMF can be considered as a supervised or unsupervised learning technique. In the case of supervised learning, the dictionary $\boldsymbol{W}$ is previously estimated from training data and matrix $\boldsymbol{H}$ only must be calculated from matrix $\boldsymbol{V}$. In the case of unsupervised learning, the two matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ must be computed jointly from $\boldsymbol{V}$. In audio applications, $\boldsymbol{V}$ is often the amplitude or power spectrogram, f denotes the frequency channel and n the time window. Figure 1.1 represents the musical score, spectrogram and unsupervised NMF of the melody of "Au clair de la lune". This figure clearly shows the interest of such a decomposition: it shows the spectra of musical notes in matrix $\boldsymbol{W}$ and their temporal activations in matrix $\boldsymbol{H}$, which makes it possible to consider both transcription and musical note separation applications.



(a) Musical score

(b) Temporal activations $\boldsymbol{H}$

(c) Spectral forms $\boldsymbol{W}$

(d) Spectrogram $\boldsymbol{V}$

Figure 1.1: Decomposition of "Au clair de la Lune" spectrogram (from [Bertin, 2009, pp. 40–41])

2 NMF theory and algorithms

Let $\mathbf{V} = [v(n,f)]_{fn}$ denote the $F \times N$ nonnegative time-frequency representation of a signal $x(t)$, where $n \in \{0,\dots,N-1\}$ is the time frame index, and $f \in \{0,\dots,F-1\}$ is the frequency index. For instance, if $\mathbf{X}$ is the $F \times N$ complex-valued *Short Time Fourier Transform* (STFT) of x, then $\mathbf{V}$ can be the magnitude spectrogram $|\mathbf{X}|$ or the power spectrogram $|\mathbf{X}|^2$ [Smaragdis and Brown, 2003]. Other choices may include perceptual frequency scales, such as constant-Q [Fuentes et al., 2013] or *Equivalent Rectangular Bandwidth* (ERB) [Vincent et al., 2010] representations.

NMF [Lee and Seung, 1999] approximates the nonnegative matrix $\mathbf{V}$ with another nonnegative matrix $\widehat{\mathbf{V}} = [\widehat{v}(n,f)]_{fn}$ with entries $v(n,f) = \sigma_x^2(n,f)$, defined as the product

$$\widehat{\mathbf{V}} = \mathbf{WH} \tag{1.1}$$

of a $F \times K$ nonnegative matrix $\mathbf{W}$ and a $K \times N$ nonnegative matrix $\mathbf{H}$ of lower rank $K < \min(F,N)$. This factorization can also be written $\widehat{\mathbf{V}} = \sum_k \widehat{\mathbf{V}}_k$, where $\widehat{\mathbf{V}}_k = [\widehat{v}_k(n,f)]_{fn} = \mathbf{w}_k\mathbf{h}_k^T$, for all $k \in \{1,\dots,K\}$, is the k-th rank-1 matrix component. The k-th column vector $\mathbf{w}_k = [w_k(f)]_f$ can be interpreted as its spectrum, and the k-th row vector $\mathbf{h}_k^T = [h_k(n)]_n$ comprises its *activation coefficients* over time. We also write $\widehat{v}_k(n,f) = w_k(f)h_k(n)$. All the parameters of the model, as well as the observed magnitude or power spectra, are elementwise nonnegative.

In this section, we first present the standard criteria for computing the NMF model parameters (Section 2.1), then we introduce probabilistic frameworks for NMF (Section 2.2), and we describe several algorithms designed for computing an NMF (Section 2.3).

2.1 Criteria for computing the NMF model parameters

Since NMF is a rank reduction technique, it involves an approximation: $\widehat{\mathbf{V}} \approx \mathbf{V}$. Computing the NMF can thus be formalized as an optimization problem: we want to minimize a measure $C(\mathbf{V} \mid \widehat{\mathbf{V}})$ of divergence between matrices $\mathbf{V}$ and $\widehat{\mathbf{V}}$. The most popular measures in the NMF literature include the squared *Euclidean* (EUC) distance [Lee and Seung, 1999], the *Kullback-Leibler* (KL) divergence [Lee and Seung, 2001], and the *Itakura-Saito* (IS) divergence [Févotte et al., 2009]. The various NMFs computed by minimizing each of these three measures are named accordingly: EUC-NMF, KL-NMF, and IS-NMF. Actually, these three measures fall under the umbrella of *β-divergences* [Nakano et al., 2010, Févotte and Idier, 2011]. Formally, they are defined for any real-valued β as

$$C^\beta(\mathbf{V} \mid \widehat{\mathbf{V}}) = \sum_{nf} d^\beta(v(n,f) \mid \widehat{v}(n,f)), \tag{1.2}$$

where

- $\forall \beta \notin \{0,1\},\ d^{\beta}(x \mid y) = \frac{1}{\beta(\beta-1)}\left(x^{\beta} + (\beta-1)y^{\beta} - \beta x y^{\beta-1}\right)$,
- $\beta = 2$ corresponds to the squared EUC distance: $d^{\mathrm{EUC}}(x \mid y) = \frac{1}{2}|x-y|^2$,
- $\beta = 1$ corresponds to the KL divergence: $d^{\mathrm{KL}}(x \mid y) = x\log(\frac{x}{y}) - x + y$,
- $\beta = 0$ corresponds to the IS divergence: $d^{\mathrm{IS}}(x \mid y) = \frac{x}{y} - \log(\frac{x}{y}) - 1$.

It can be easily proved that $\forall x > 0$, the function $y \mapsto d^{\beta}(x \mid y)$ is convex with respect to y if and only if $\beta \in [1,2]$ [Févotte and Idier, 2011]. This means that minimizing $C^{\beta}(\mathbf{V} \mid \mathbf{WH})$ with respect to $\mathbf{H}$ with $\mathbf{W}$ fixed, or conversely with respect to $\mathbf{W}$ with $\mathbf{H}$ fixed, is a convex optimization problem if and only if $\beta \in [1,2]$. This convexity property is particularly convenient in a pretrained framework, where matrix $\mathbf{W}$ is fixed, and only matrix $\mathbf{H}$ is estimated from the observed data. Optimization algorithms are then insensitive to initialization, which might explain the better success of KL-NMF and EUC-NMF compared with IS-NMF in the NMF literature.

However, in the context of learning-free separation, whatever the value of β, minimizing $C^{\beta}(\mathbf{V} \mid \mathbf{WH})$ jointly with respect to $\mathbf{W}$ and $\mathbf{H}$ is not a convex optimization problem. Indeed, this factorization is not unique, since we also have $\widehat{\mathbf{V}} = \mathbf{W}'\mathbf{H}'$ with $\mathbf{W}' = \mathbf{W}\mathbf{\Lambda}\mathbf{\Pi}$ and $\mathbf{H}' = \mathbf{\Pi}^T\mathbf{\Lambda}^{-1}\mathbf{H}$, where $\mathbf{\Lambda}$ can be any $K \times K$ diagonal matrix with positive diagonal entries, and $\mathbf{\Pi}$ can be any $K \times K$ permutation matrix. Note that the nonuniqueness of the model is actually ubiquitous in source separation and is generally not considered as a problem: sources are recovered up to a scale factor and a permutation. In the case of NMF however, other kinds of indeterminacies may also exist [Laurberg et al., 2008]. Due to the existence of local minima, optimization algorithms become sensitive to initialization (cf. Section 2.3). In practice, with a random initialization, there is no longer any guarantee to converge to a solution that is helpful for source separation. For this reason, advanced NMF models have later been proposed to enforce some specific desired properties in the decomposition (cf. Section 3).

2.2 Probabilistic frameworks for NMF

Computing an NMF can also be formalized as a parametric estimation problem based on a probabilistic model, that involves both observed and latent (hidden) random variables. Typically, observed variables are related to matrix $\mathbf{V}$, whereas latent variables are related to matrices $\mathbf{W}$ and $\mathbf{H}$. The main advantages of using a probabilistic framework are the facility of exploiting some a priori knowledge that we may have about matrices $\mathbf{W}$ and $\mathbf{H}$, and the existence of well-known statistical inference techniques, such as the *Expectation-Maximization* (EM) algorithm.

Popular probabilistic models of nonnegative time-frequency representations include Gaussian models that are equivalent to EUC-NMF [Schmidt and Laurberg, 2008] (Section 2.2.1), and count models, such as the celebrated *Probabilistic Latent Component Analysis* (PLCA) [Shashanka et al., 2008] (Section 2.2.2) and the Poisson NMF model based on the Poisson distribution [Virtanen et al., 2008] (Section 2.2.3), that are both related to KL-NMF. However these probabilistic models do not account for the fact that matrix $\mathbf{V}$ has been generated from a time-domain signal $x(t)$. As a result, they can be used to estimate a nonnegative time-frequency representation, but they are not able to account for the phase, that is necessary to reconstruct a time-domain signal.

Other probabilistic frameworks focus on power or magnitude spectrograms, and intend to directly model the STFT $\mathbf{X}$ instead of the nonnegative time-frequency representation $\mathbf{V}$, in order to permit the resynthesis of time-domain signals. The main advantage of this approach is the ability to account for the phase and, in source separation applications, to provide a theoretical ground for using time-frequency masking techniques. Such models include Gaussian models that are equivalent to IS-NMF [Févotte et al., 2009] (Section 2.2.4) and the Cauchy NMF model based on the Cauchy distribution [Liutkus et al., 2015], that both fall under the umbrella of α-stable models [Liutkus and Badeau, 2015] (Section 2.2.5).

2.2.1 Gaussian noise model

A simple probabilistic model for EUC-NMF was presented by Schmidt and Laurberg [2008]: $\mathbf{V} = \mathbf{WH} + \mathbf{U}$, where matrices $\mathbf{W}$ and $\mathbf{H}$ are seen as deterministic parameters, and the entries of matrix $\mathbf{U}$ are Gaussian, independent and identically distributed (i.i.d.): $u(n,f) \sim \mathcal{N}(u(n,f) \mid 0, \sigma_u^2)$. Then the log-likelihood of matrix $\mathbf{V}$ is $\log p(\mathbf{V} \mid$

$\mathbf{W},\mathbf{H}) = -\frac{1}{2\sigma_u^2}\|\mathbf{V}-\mathbf{WH}\|_F^2 + \text{cst} = -\frac{1}{\sigma_u^2}C^2(\mathbf{V} \mid \widehat{\mathbf{V}}) + \text{cst}$ where cst denotes a constant additive term, as defined in (1.2) with $\beta = 2$. Therefore *Maxmimum Likelihood* (ML) estimation of $(\mathbf{W},\mathbf{H})$ is equivalent to EUC-NMF. The main drawback of this generative model is that it does not enforce the nonnegativity of $\mathbf{V}$, whose entries might take negative values.

2.2.2 Probabilistic latent component analysis

PLCA [Shashanka et al., 2008] is a count model that views matrix $\widehat{\mathbf{V}}$ as a probability distribution (normalized so that $\sum_{nf}\widehat{v}(n,f) = 1$). The observation model is the following one: the probability distribution $P(n,f) = \widehat{v}(n,f)$ is sampled M times to produce M independent time-frequency pairs (n_m, f_m), $m \in \{1,\dots,M\}$. Then matrix $\mathbf{V}$ is generated as a histogram: $v(n,f) = \frac{1}{M}\sum_m \delta_{(n_m,f_m)}(n,f)$, that also satisfies $\sum_{nf}\widehat{z}(n,f) = 1$. The connection with NMF is established by introducing a latent variable k that is also sampled M times to produce k_m, $m \in \{1,\dots,M\}$. More precisely, it is assumed that (k_m, n_m) are first sampled together according to distribution $P(k,n) = h_k(n)$, and that f_m is then sampled given k_m according to distribution $P(f \mid k) = w_k(f)$, resulting in the joint distribution $P(n,f,k) = P(k,n)P(f \mid k) = \widehat{v}_k(n,f)$. Then $P(n,f)$ is the marginal distribution resulting from the joint distribution $P(n,f,k)$: $P(n,f) = \sum_k P(n,f,k) = \widehat{v}(n,f)$. Finally, note that another convenient formulation of PLCA is to simply state that $\mathbf{v}(n) \sim \mathcal{M}(\mathbf{v}(n) \mid \|\mathbf{v}(n)\|_1, \mathbf{Wh}(n))$ where $\mathbf{v}(n)$ and $\mathbf{h}(n)$ are the n-th columns of matrices $\mathbf{V}$ and $\mathbf{H}$, respectively, $\mathcal{M}$ denotes the multinomial distribution, and $\mathbf{h}(n)$ and the columns of matrix $\mathbf{W}$ are vectors that sum to 1.

In Section 2.3, it will be shown that this probabilistic model is closely related to KL-NMF. Indeed, the update rules obtained by applying the EM algorithm are formally equivalent to KL-NMF multiplicative update rules (cf. Section 2.3.3).

2.2.3 Poisson NMF model

The Poisson NMF model [Virtanen et al., 2008] is another count model, that assumes that the observed nonnegative matrix $\mathbf{V}$ is generated as the sum of K independent, nonnegative latent components $\mathbf{V}_k$. The entries $v_k(n,f)$ of matrix $\mathbf{V}_k$ are assumed independent and *Poisson*-distributed: $v_k(n,f) \sim \mathcal{P}(v_k(n,f) \mid \widehat{v}_k(n,f))$. The Poisson distribution is defined for any positive integer $\widehat{v}$ as $\mathcal{P}(\widehat{v} \mid \lambda) = \frac{e^{-\lambda}\lambda^{\widehat{v}}}{\widehat{v}!}$, where λ is the intensity parameter and $\widehat{v}!$ is the factorial of $\widehat{v}$. A nice feature of the Poisson distribution is that the sum of K independent Poisson random variables with intensity parameters λ_k is a Poisson random variable with intensity parameter $\lambda = \sum_k \lambda_k$. Consequently, $v(n,f) \sim \mathcal{P}(v(n,f) \mid \widehat{v}(n,f))$. The NMF model $\widehat{\mathbf{V}} = \mathbf{WH}$ can thus be computed by maximizing $P(\mathbf{V} \mid \mathbf{W},\mathbf{H}) = \prod_{nf}\mathcal{P}(v(n,f) \mid \widehat{v}(n,f))$. It can be noticed that $\log P(\mathbf{V} \mid \mathbf{W},\mathbf{H}) = -C^1(\mathbf{V} \mid \widehat{\mathbf{V}})$, as defined in (1.2) with $\beta = 1$. Therefore ML estimation of $(\mathbf{W},\mathbf{H})$ is equivalent to KL-NMF.

2.2.4 Gaussian composite model

The Gaussian *composite model*introduced by Févotte et al. [2009] exploits a feature of the Gaussian distribution that is similar to that of the Poisson distribution: a sum of K independent Gaussian random variables of means μ_k and variances σ_k^2 is a Gaussian random variable of mean $\mu = \sum_k \mu_k$ and variance $\sigma^2 = \sum_k \sigma_k^2$. The main difference with the Poisson NMF model is that, instead of modeling the nonnegative time-frequency representation $\mathbf{V}$, IS-NMF aims to model the complex STFT $\mathbf{X}$, such that $\mathbf{V} = |\mathbf{X}|^2$. The observed complex matrix $\mathbf{X}$ is thus generated as the sum of K independent complex latent components $\mathbf{X}_k$ (cf. Fig 1.2). The entries of matrix $\mathbf{X}_k$ are assumed independent and complex Gaussian distributed: $x_k(n,f) \sim \mathcal{N}_c(x_k(n,f) \mid 0, \widehat{v}_k(n,f))$. Here the complex Gaussian distribution is defined as $\mathcal{N}_c(x \mid \mu,\sigma^2) = \frac{1}{\pi\sigma^2}\exp(-\frac{|x-\mu|^2}{\sigma^2})$, where μ and σ^2 are the mean and variance parameters. Consequently, $x(n,f) = \sum_k x_k(n,f) \sim \mathcal{N}_c(x(n,f) \mid 0, \widehat{v}(n,f))$. The NMF model $\widehat{\mathbf{V}} = \mathbf{WH}$ can thus be computed by maximizing $p(\mathbf{X} \mid \mathbf{W},\mathbf{H}) = \prod_{nf}\mathcal{N}_c(x(n,f) \mid 0, \widehat{v}(n,f))$. It can be noticed that $\log p(\mathbf{X} \mid \mathbf{W},\mathbf{H}) = -C^0(\mathbf{V} \mid \widehat{\mathbf{V}})$, as defined in (1.2) with $\beta = 0$. Therefore ML estimation of $(\mathbf{W},\mathbf{H})$ is equivalent to IS-NMF.

In a source separation application, the main practical advantage of this Gaussian composite model is that the *Minimum Mean Square Error* (MMSE) estimates of the sources are obtained by time-frequency masking, in a way that is closely related to Wiener filtering. Indeed, suppose now that the observed signal $x(t)$ is the sum of J unknown source signals $s_j(t)$, so that $x(n,f) = \sum_j s_j(n,f)$, and that each source follows an IS-NMF model:

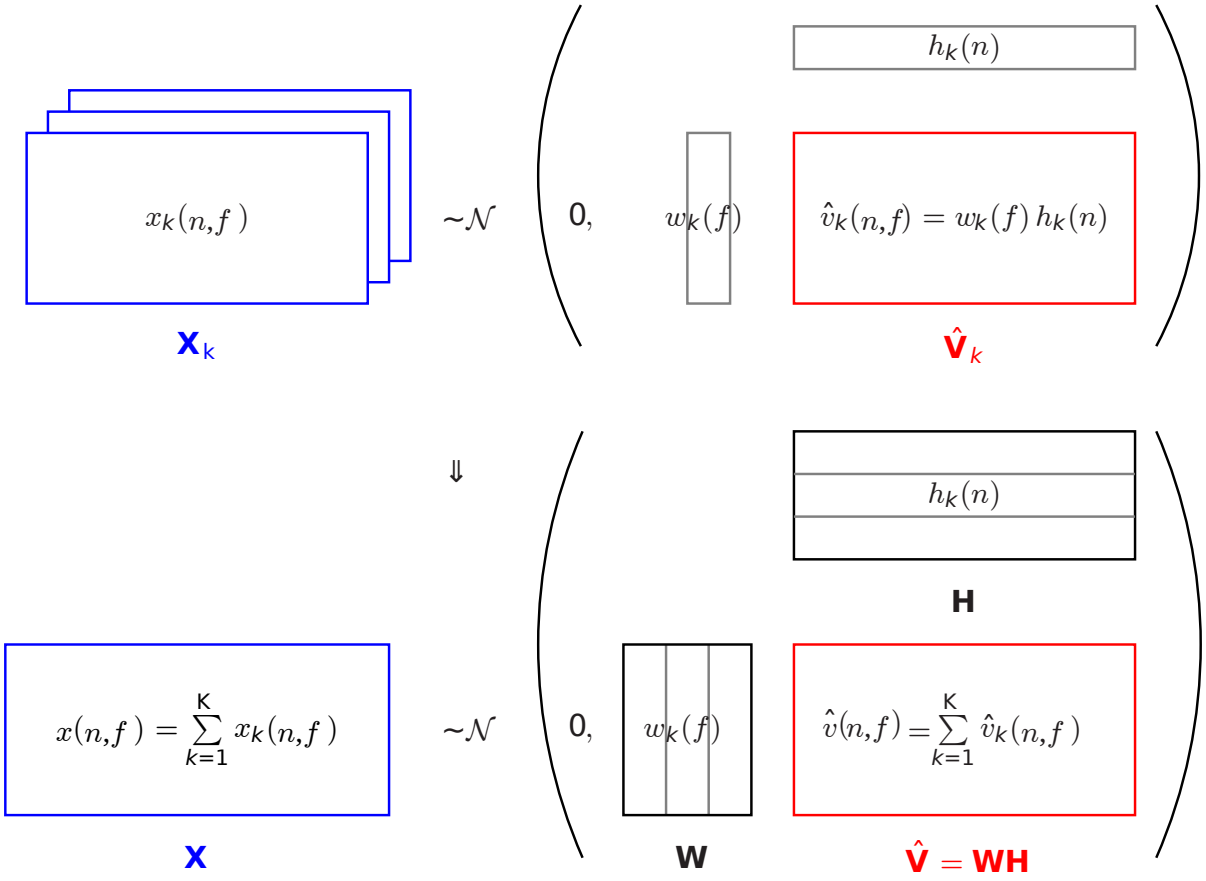


Figure 1.2: Gaussian composite model (IS-NMF) by Févotte et al. [2009]

$s_j(n,f) \sim \mathcal{N}_c(s_j(n,f) \mid 0, \widehat{v}_j(n,f))$ where $\widehat{v}_j(n,f) = \sigma^2_{s_j}(n,f)$ denotes the entries of matrix $\widehat{\mathbf{V}}_j = \mathbf{W}_j\mathbf{H}_j$. Then the minimum of the mean square error (MSE) criterion $\sum_{nf} \mathbb{E}\{|s_j(n,f) - \widehat{s}_j(n,f)|^2 \mid x(n,f)\}$ is reached when

$$\forall n,f,\ \widehat{s}_j(n,f) = \mathbb{E}\{s_j(n,f) \mid x(n,f)\} = m_j(n,f)x(n,f), \tag{1.3}$$

where the time-frequency mask $m_j(n,f)$ is defined as

$$m_j(n,f) = \frac{\widehat{v}_j(n,f)}{\sum_{j'} \widehat{v}_{j'}(n,f)}. \tag{1.4}$$

2.2.5 α-stable NMF models

Despite its nice features, the Gaussian model introduced in the previous section presents two drawbacks. Firstly, it amounts to assuming the additivity of the source power spectrograms, whereas several experimental studies have shown that the additivity of magnitude spectrograms is a better fit (see Liutkus and Badeau [2015] and references therein). Secondly, the IS divergence is not convex, which leads to increased optimization issues due to the existence of local minima. In order to circumvent these problems, a generalization of this model was introduced by Liutkus and Badeau [2015], based on isotropic complex α-*stable* distributions denoted $S\alpha S_c$, which stands for complex symmetric α-stable. This is a family of heavy-tailed probability distributions defined for any $\alpha \in]0,2]$, which do not have a closed-form expression, except in the particular cases $\alpha = 2$, which corresponds to the complex

Gaussian distribution, and $\alpha = 1$, which corresponds to the isotropic complex *Cauchy* distribution. In the general case, the distribution is defined by its characteristic function: $x \sim S\alpha S_c(x \mid \sigma) \Leftrightarrow \phi_z(\theta) = \mathbb{E}\{e^{j\Re(\theta^* x)}\} = e^{-\sigma^\alpha |\theta|^\alpha}$ for any complex-valued θ, where $\sigma > 0$ is the scale parameter (which corresponds to the standard deviation in the Gaussian case). These probability distributions enjoy the same nice feature shared by the Poisson and Gaussian distributions: a sum of K independent isotropic complex α-stable random variables of scale parameters σ_k is an isotropic complex α-stable random variable of scale parameter $\sigma^\alpha = \sum_k \sigma_k^\alpha$.

In this context, the observed STFT matrix $\mathbf{X}$ is again modeled as the sum of K independent latent components $\mathbf{X}_k$. The entries of matrix $\mathbf{X}_k$ are independent and isotropic complex α-stable: $x_k(n,f) \sim S\alpha S_c(x_k(n,f) \mid \sigma_k(n,f))$, where $\widehat{v}_k(n,f) = \sigma_k^\alpha(n,f)$ is called an α-spectrogram. Thus $x(n,f) = \sum_k x_k(n,f) \sim S\alpha S_c(x(n,f) \mid \sigma(n,f))$, with

$$\widehat{v}(n,f) = \sigma^\alpha(n,f) = \sum_k \sigma_k^\alpha(n,f) = \sum_k \widehat{v}_k(n,f). \tag{1.5}$$

When the distribution has a closed-form expression (i.e., $\alpha = 1$ or 2), the NMF model $\widehat{\mathbf{V}} = \mathbf{WH}$ can still be estimated in the ML sense, otherwise different inference methods are required. In the Cauchy case [Liutkus et al., 2015], it has been experimentally observed that Cauchy NMF is much less sensitive to initialization than IS-NMF and produces meaningful basis spectra for source separation.

In a source separation application, we again suppose that the observed signal $x(t)$ is the sum of J unknown source signals $s_j(t)$, so that $x(n,f) = \sum_j s_j(n,f)$, and that each source follows an isotropic complex α-stable NMF model: $s_j(n,f) \sim S\alpha S_c(s_j(n,f) \mid \widehat{v}_j^{1/\alpha}(n,f))$ where $\widehat{v}_j(n,f)$ denotes the entries of matrix $\widehat{\mathbf{V}}_j = \mathbf{W}_j\mathbf{H}_j$. Then the MSE criterion is no longer defined for all $\alpha \in (0,2]$, but in any case, the posterior mean $\widehat{s}_j(n,f) = \mathbb{E}\{s_j(n,f) \mid x(n,f)\}$ is still well-defined, and admits the same expression as in (1.3) and (1.4).

2.2.6 Choosing a particular NMF model

When choosing a particular NMF model for a given source separation application, several criteria may be considered, including the following ones:

Robustness to initialization: Cauchy NMF has proved to be more robust to initialization than all other probabilistic NMF models. Besides, in the context of pretrained source separation, the Gaussian noise model (related to EUC-NMF) and the PLCA/Poisson NMF models (related to KL-NMF) lead to a convex optimization problem with a unique minimum, which is not the case of the Gaussian composite model (related to IS-NMF).

Source reconstruction: only the α-stable NMF models, including IS-NMF and Cauchy NMF, provide a theoretical ground for using Wiener filtering in order to reconstruct time-domain signals.

Existence of closed-form update rules: ML estimation of the model parameters is tractable for all NMF models except α-stable models with $\alpha \neq 1, 2$.

2.3 Algorithms for NMF

In the literature, various algorithms have been designed for computing an NMF, including the famous multiplicative update rules [Lee and Seung, 2001], the alternated least squares method [Finesso and Spreij, 2004], and the projected gradient method [Lin, 2007]. In Section 2.3.1, we present the multiplicative update rules, that form the most celebrated NMF algorithm, and we summarize their convergence properties. Then in the following sections, we present some algorithms dedicated to the probabilistic frameworks introduced in Section 2.2.

2.3.1 Multiplicative update rules

The basic idea of *multiplicative update* rules is that the nonnegativity constraint can be easily enforced by updating the previous values of the model parameters by multiplication with a nonnegative scale factor. A heuristic way of deriving these updates consists in decomposing the gradient of the cost function $C(\mathbf{V} \mid \widehat{\mathbf{V}})$, e.g., the β-divergence introduced in (1.2), as the difference of two nonnegative terms: $\nabla_{\mathbf{W}} C(\mathbf{V} \mid \widehat{\mathbf{V}}) = \nabla^+_{\mathbf{W}} C(\mathbf{V} \mid \widehat{\mathbf{V}}) - \nabla^-_{\mathbf{W}} C(\mathbf{V} \mid \widehat{\mathbf{V}})$, where

$\nabla^+_{\mathbf{W}} C(\mathbf{V} \mid \widehat{\mathbf{V}}) \geq 0$ and $\nabla^-_{\mathbf{W}} C(\mathbf{V} \mid \widehat{\mathbf{V}}) \geq 0$, meaning that all the entries of these two matrices are nonnegative. Then matrix $\mathbf{W}$ can be updated as $\mathbf{W} \leftarrow \mathbf{W} \circ (\nabla^-_{\mathbf{W}} C(\mathbf{V} \mid \widehat{\mathbf{V}}) / \nabla^+_{\mathbf{W}} C(\mathbf{V} \mid \widehat{\mathbf{V}}))^\eta$, where $\circ$ denotes elementwise matrix product, / denotes elementwise matrix division, the matrix exponentiation must be understood elementwise, and $\eta > 0$ is a stepsize similar to that involved in a gradient descent [Badeau et al., 2010]. The same update can be derived for matrix $\mathbf{H}$, and then matrices $\mathbf{W}$ and $\mathbf{H}$ can be updated in turn, until convergence[1]. Note that the decomposition of the gradient as a difference of two nonnegative terms is not unique, and different choices can be made, leading to different multiplicative update rules. In the case of the β-divergence, the standard multiplicative update rules are expressed as follows [Févotte and Idier, 2011]:

$$\mathbf{W} \leftarrow \mathbf{W} \circ \left(\frac{(\mathbf{V} \circ (\mathbf{WH})^{\beta-2})\mathbf{H}^T}{(\mathbf{WH})^{\beta-1}\mathbf{H}^T} \right)^\eta \tag{1.6}$$

$$\mathbf{H} \leftarrow \mathbf{H} \circ \left(\frac{\mathbf{W}^T(\mathbf{V} \circ (\mathbf{WH})^{\beta-2})}{\mathbf{W}^T(\mathbf{WH})^{\beta-1}} \right)^\eta \tag{1.7}$$

where matrix division and exponentiation must be understood elementwise. By using the auxiliary function approach, Nakano et al. [2010] proved that the cost function $C^\beta(\mathbf{V} \mid \widehat{\mathbf{V}})$ is nonincreasing under these updates when the stepsize η is given by $\eta = \frac{1}{2-\beta}$ for $\beta < 1$, $\eta = 1$ for $1 \leq \beta \leq 2$, and $\eta = \frac{1}{\beta-1}$ for $\beta > 2$. In addition, Févotte and Idier [2011] proved that the same cost function is nonincreasing under (1.6)–(1.7) for $\eta = 1$ and for all $\beta \in [0, 2]$ (which includes the popular EUC, KL and IS-NMF). They also proved that these updates correspond to a *Majorization-Minimization* (MM) algorithm when the stepsize η is expressed as a given function of β, which is equal to 1 for all $\beta \in [1, 2]$, and they correspond to a majorization-equalization algorithm for $\eta = 1$ and $\beta = 0$. However, contrary to a widespread belief [Lee and Seung, 2001], the decrease of the cost function is not sufficient to prove the convergence of the algorithm to a local or global minimum. Badeau et al. [2010] analyzed the convergence of multiplicative update rules by means of Lyapunov's stability theory. In particular, it was proved that:

- There is $\eta^{\max} > 0$ such that these rules are exponentially or asymptotically stable for all $\eta \in (0, \eta^{\max})$. Moreover, $\forall \beta$, the upper bound $\eta^{\max}$ is such that $\eta^{\max} \in (0, 2]$, and if $\beta \in [1, 2]$, $\eta^{\max} = 2$.
- These rules are unstable if $\eta \notin [0, 2]$, $\forall \beta$.

In practice, the step size η permits us to control the convergence rate of the algorithm.

Note that, due to the nonuniqueness of NMF, there is a scaling and permutation ambiguity between matrices $\mathbf{W}$ and $\mathbf{H}$ (cf. Section 2.1). Therefore, when $\mathbf{W}$ and $\mathbf{H}$ are to be updated in turn, numerical stability can be improved by renormalizing the columns of $\mathbf{W}$ (resp. the rows of $\mathbf{H}$), and scaling the rows of $\mathbf{H}$ (resp. the columns of $\mathbf{W}$) accordingly, so as to keep the product $\mathbf{WH}$ unchanged.

Finally, a well-known drawback of most NMF algorithms is the sensitivity to initialization, that is due to the multiplicity of local minima of the cost function (cf. Section 2.1). Many initialization strategies were thus proposed in the literature [Cichocki et al., 2009]. In the case of IS-NMF multiplicative update rules, a *tempering* approach was proposed by Bertin et al. [2009]. The basic idea is the following one: since the β-divergence is convex for all $\beta \in [1, 2]$, but not for $\beta = 0$, the number of local minima is expected to increase when β goes from 2 to 0. Therefore a simple solution for improving the robustness to initialization consists in making parameter β vary from 2 to 0 over the iterations of the algorithm. Nevertheless, the best way of improving the robustness to initialization in general is to select a robust NMF criterion, such as that involved in Cauchy NMF (cf. Section 2.2.5).

2.3.2 The EM algorithm and its variants

As mentioned in Section 2.2, one advantage of using a probabilistic framework for NMF is the availability of classical inference techniques, whose convergence properties are well-known. Classical algorithms used in the NMF literature include the EM algorithm [Shashanka et al., 2008], the space-alternating generalized EM algorithm [Févotte et al., 2009], variational Bayesian (VB) inference [Badeau and Drémeau, 2013], and *Markov chain Monte Carlo* [Simsekli and Cemgil, 2012].

[1] This iterative algorithm can stop, e.g., when the decrease of the β-divergence, or when the distance between the successive iterates of matrices $\mathbf{W}$ and $\mathbf{H}$, goes below a given threshold.

Below, we introduce the basic principles of the space-alternating generalized EM algorithm [Fessler and Hero, 1994], which includes the regular EM algorithm as a particular case. We then apply the EM algorithm to the PLCA framework described in Section 2.2.2, and the space-alternating generalized EM algorithm to the Gaussian composite model described in Section 2.2.4.

Consider a random observed dataset $\mathcal{X}$, whose probability distribution is parameterized by a parameter set $\boldsymbol{\theta}$, that is partitioned as $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}_k$. The space-alternating generalized EM algorithm aims to estimate parameters $\boldsymbol{\theta}_k$ iteratively, while guaranteeing that the likelihood $p(\mathcal{X} \mid \boldsymbol{\theta})$ is nondecreasing over the iterations. It requires choosing for each subset $\boldsymbol{\theta}_k$ a *hidden data* space which is complete for this particular subset, i.e., a latent dataset $\mathcal{X}_k$ such that $p(\mathcal{X}, \mathcal{X}_k \mid \boldsymbol{\theta}) = p(\mathcal{X} \mid \mathcal{X}_k, \{\boldsymbol{\theta}_{k'}\}_{k' \neq k}) p(\mathcal{X}_k \mid \boldsymbol{\theta})$. The algorithm iterates over both the iteration index and over k. For each iteration and each k, it is composed of an expectation step (*E-step*) and a maximization step (*M-step*):

- E-step: evaluate $\mathcal{Q}_k(\boldsymbol{\theta}_k) = \mathbb{E}\{\log p(\mathcal{X}_k \mid \boldsymbol{\theta}_k, \{\boldsymbol{\theta}_{k'}\}_{k' \neq k}) \mid \mathcal{X}, \boldsymbol{\theta}\}$;
- M-step: compute $\boldsymbol{\theta}_k = \operatorname{argmax}_{\boldsymbol{\theta}_k} \mathcal{Q}_k(\boldsymbol{\theta}_k)$.

The regular EM algorithm corresponds to the particular case $K = 1$, where $\mathcal{X}$ is a deterministic function of the complete data space.

2.3.3 Application of the EM algorithm to PLCA

Shashanka et al. [2008] applied the EM algorithm to the PLCA model described in Section 2.2.2. The observed dataset is $\mathcal{X} = \{n_m, f_m\}_m$, the parameter set is $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$, and the complete data space is $\{n_m, f_m, k_m\}_m$. Then:

- The E-step consists in computing $P(k \mid n, f) = \frac{P(n,f,k)}{\sum_{k'} P(n,f,k')} = \frac{\widehat{v}_k(n,f)}{\widehat{v}(n,f)}$, that appears in the expression $\mathcal{Q}(\boldsymbol{\theta}) = \sum_{nf} v(n,f) \sum_k P(k \mid n, f) \log\left(w_k(f) h_k(n)\right)$.
- The M-step consists in maximizing $\mathcal{Q}(\boldsymbol{\theta})$ with respect to $w_k(f)$ and $h_k(n)$, subject to $\forall k$, $\sum_f w_k(f) = 1$ and $\sum_{kn} h_k(n) = 1$. Given that $\sum_{nf} v(n,f) = 1$, we get:

$$h_k(n) \leftarrow \frac{\sum_f v(n,f) P(k \mid n, f)}{\sum_{k'n'f} v(n',f) P(k' \mid n', f)} = h_k(n) \sum_f w_k(f) \frac{v(n,f)}{\widehat{v}(n,f)}, \tag{1.8}$$

$$w_k(f) \leftarrow \frac{\sum_n v(n,f) P(k \mid n, f)}{\sum_{nf'} v(n,f') P(k \mid n, f')} = \frac{\tilde{w}_k(f)}{\sum_{f'} \tilde{w}_k(f')}, \tag{1.9}$$

  where $\tilde{w}_k(f) = w_k(f) \sum_n h_k(n) \frac{v(n,f)}{\widehat{v}(n,f)}$.

It is easy to check that this algorithm is identical to the multiplicative update rules for KL divergence, as described in (1.6)–(1.7) with $\eta = \beta = 1$, up to a scaling factor in $\mathbf{H}$ due to the normalization of matrix $\mathbf{V}$ [Shashanka et al., 2008].

2.3.4 Application of the space-alternating generalized EM algorithm to the Gaussian composite model

Févotte et al. [2009] applied the *Space Alternating Generalized EM* (SAGE) algorithm to the Gaussian composite model described in Section 2.2.4. The observed dataset is $\mathcal{X} = \mathbf{X}$, the k-th parameter set is $\boldsymbol{\theta}_k = \{\mathbf{w}_k, \mathbf{h}_k\}$, and the k-th complete latent dataset is $\mathcal{X}_k = \mathbf{X}_k$. Then:

- The E-step consists in computing $\mathbf{V}_k = \frac{\widehat{\mathbf{V}}_k^2}{\widehat{\mathbf{V}}^2} \circ \mathbf{V} + \frac{\widehat{\mathbf{V}}_k \circ (\widehat{\mathbf{V}} - \widehat{\mathbf{V}}_k)}{\widehat{\mathbf{V}}}$, that appears in the expression $\mathcal{Q}_k(\boldsymbol{\theta}_k) = -C^0(\mathbf{V}_k \mid \widehat{\mathbf{V}}_k)$, where criterion C^0 was defined in (1.2) with $\beta = 0$.
- The M-step computes $h_k(n) \leftarrow \frac{1}{F} \sum_f \frac{v_k(n,f)}{w_k(f)}$ and $w_k(f) \leftarrow \frac{1}{N} \sum_n \frac{v_k(n,f)}{h_k(n)}$.

Note that it has been experimentally observed by Févotte et al. [2009] that this space-alternating generalized EM algorithm converges more slowly than the IS-NMF multiplicative update rules described in (1.6)–(1.7) for $\eta = 1$ and $\beta = 0$.

3 Advanced NMF models

The basic NMF model presented in Section 1 has proved successful for addressing a variety of audio source separation problems. Nevertheless, the source separation performance can still be improved by exploiting prior knowledge that we may have about the source signals. For instance, we know that musical notes and voiced sounds have a harmonic spectrum (or, more generally, an inharmonic or a sparse spectrum), and that both their spectral envelope and their temporal power profile have smooth variations. On the opposite, percussive sounds rather have a smooth spectrum, and a sparse temporal power profile. It may thus be desirable to impose properties such as *harmonicity*, *smoothness*, and *sparsity* on either the spectral matrix $\mathbf{W}$ or the activation matrix $\mathbf{H}$ in the NMF $\widehat{\mathbf{V}} = \mathbf{WH}$. For that purpose, it is possible to apply either hard constraints, e.g., by parameterizing matrix $\mathbf{W}$ or $\mathbf{H}$, or soft constraints, e.g., by adding a regularization term to the criterion (1.2), or by introducing the prior distributions of $\mathbf{W}$ or $\mathbf{H}$ in the probabilistic frameworks introduced in Section 2.2 (Bayesian approach). Examples of such regularizations are described in Section 3.1. Note that another possible way of exploiting prior information is to use a predefined dictionary $\mathbf{W}$ trained on a training dataset.

In other respects, audio signals are known to be nonstationary, therefore it is useful to consider that some characteristics such as the fundamental frequency or the spectral envelope may vary over time. Such nonstationary models will be presented in Section 3.2.

3.1 Regularizations

In this section, we present a few examples of NMF regularizations, including sparsity (Section 3.1.1), group-sparsity (Section 3.1.2), harmonicity and spectral smoothness (Section 3.1.3), and inharmonicity (Section 3.1.4).

3.1.1 Sparsity

Since NMF is well suited to the problem of separating audio signals formed of a few repeated audio events, it is often desirable to enforce the sparsity of matrix $\mathbf{H}$.

The most straightforward way of doing so is to add to the NMF criterion a sparsity-promoting regularization term. Ideally, sparsity is measured by the ℓ_0 norm, which counts the number of nonzero entries in a vector. However, optimizing a criterion involving the ℓ_0 norm raises intractable combinatorial issues. In the optimization literature, the ℓ_1 norm is often preferred, because it is the tightest convex relaxation of the ℓ_0 norm. Therefore the criterion $C^\beta(\mathbf{V} \mid \widehat{\mathbf{V}})$ in (1.2) may be replaced with

$$C(\mathbf{V} \mid \widehat{\mathbf{V}}) = C^\beta(\mathbf{V} \mid \widehat{\mathbf{V}}) + \lambda \sum_k \|\mathbf{h}_k\|_1, \tag{1.10}$$

where $\lambda > 0$ is a tradeoff parameter to be tuned manually, as suggested, e.g., by Hurmalainen et al. [2015].

However, if the NMF is embedded in a probabilistic framework such as those introduced in Section 2.2, sparsity is rather enforced by introducing an appropriate prior distribution of matrix $\mathbf{H}$. In this case, $\mathbf{H}$ is estimated by maximizing its posterior probability given $\mathbf{V}$, or equivalently the *Maximum a Posteriori* (MAP) criterion $\log p(\mathbf{V} \mid \mathbf{W}, \mathbf{H}) + \log p(\mathbf{H})$, instead of the log-likelihood $\log p(\mathbf{V} \mid \mathbf{W}, \mathbf{H})$. For instance, Kameoka et al. [2009] consider a generative model similar to the Gaussian noise model presented in Section 2.2.1, where the sparsity of matrix $\mathbf{H}$ is enforced by means of a generalized Gaussian prior:

$$p(\mathbf{H}) = \prod_{kn} \frac{1}{2\Gamma\left(1 + \frac{1}{p}\right)\sigma} e^{-\frac{|h_k(n)|^p}{\sigma^p}}, \tag{1.11}$$

where $\Gamma(\cdot)$ denotes the gamma function, parameter p promotes sparsity if $0 < p < 2$, and the case $p = 2$ corresponds to the standard Gaussian distribution.

In the PLCA framework described in Section 2.2.2, the entries of $\mathbf{H}$ are the discrete probabilities $P(k, n)$. By noticing that the entropy $\mathbb{H}\{\mathbf{H}\}$ of this discrete probability distribution is related to the sparsity of matrix $\mathbf{H}$ (the lower $\mathbb{H}\{\mathbf{H}\}$, the sparser $\mathbf{H}$), a suitable sparsity-promoting prior is the so-called *entropic prior* [Shashanka et al., 2008], defined as $p(\mathbf{H}) \propto e^{-\beta\mathbb{H}\{\mathbf{H}\}}$, where $\beta > 0$.

3.1.2 Group sparsity

Now suppose that the observed signal $x(t)$ is the sum of J unknown source signals $s_j(t)$ for $j \in \{1, \ldots, J\}$, whose spectrograms $\mathbf{V}_j$ are approximated as $\widehat{\mathbf{V}}_j = \mathbf{W}_j\mathbf{H}_j$, as in Section 2.2.4. Then the spectrogram $\mathbf{V}$ of $x(t)$ is approximated with the NMF $\sum_j \widehat{\mathbf{V}}_j = \mathbf{WH}$, where $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_J]$ and $\mathbf{H} = [\mathbf{H}_1^T, \ldots, \mathbf{H}_J^T]^T$. In this context, it is natural to expect that if a given source j is inactive at time n, then all the entries in the n-th column of $\mathbf{H}_j$ are zero. Such a property can be enforced by using *group sparsity*. A well-known group sparsity regularization term is the mixed ℓ_2-ℓ_1 norm: $\|\mathbf{H}\|_{2,1} = \sum_{jn} \|\mathbf{h}_j(n)\|_2$ where $\mathbf{h}_j(n)$ is the n-th column of matrix $\mathbf{H}_j$, as suggested, e.g., by Hurmalainen et al. [2015]. Indeed, the minimization of the criterion (1.10) involving this regularization term tends to enforce sparsity over both n and j, while ensuring that the whole vector $\mathbf{h}_j(n)$ gets close to zero for most values of n and j.

Lefevre et al. [2011] proposed a group sparsity prior for the IS-NMF probabilistic framework described in Section 2.2.4. The idea is to consider a prior distribution of matrix $\mathbf{H}$ such that all vectors $\mathbf{h}_j(n)$ are independent: $p(\mathbf{H}) = \prod_{jn} p(\mathbf{h}_j(n))$. Each $p(\mathbf{h}_j(n))$ is chosen so as to promote near-zero vectors. Then, as in Section 3.1.1, the NMF parameters are estimated in the MAP sense: $(\mathbf{W}, \mathbf{H}) = \operatorname{argmax}_{\mathbf{W},\mathbf{H}} \log p(\mathbf{X} \mid \mathbf{W}, \mathbf{H}) + \sum_{jn} \log p(\mathbf{h}_j(n))$, where $p(\mathbf{X} \mid \mathbf{W}, \mathbf{H})$ was defined in Section 2.2.4.

3.1.3 Harmonicity and spectral smoothness

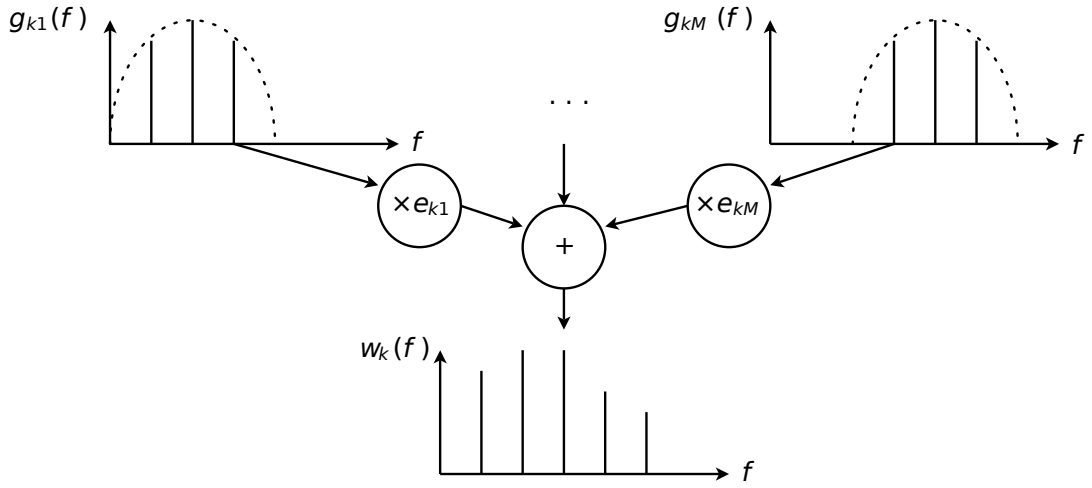


Figure 1.3: Harmonic NMF model by Vincent et al. [2010] and Bertin et al. [2010]

Contrary to sparsity, harmonicity in matrix $\mathbf{W}$ is generally enforced as a hard constraint, by using parametric models, whose parameter set includes the fundamental frequency. For instance, Vincent et al. [2010] and Bertin et al. [2010] parameterized the spectrum vector $\mathbf{w}_k$ as a nonnegative linear combination of M narrowband, harmonic spectral patterns (cf. Fig. 1.3): $w_k(f) = \sum_m e_{km} g_{km}(f)$, where all spectral patterns $\{g_{km}(f)\}_m$ share the same fundamental frequency $\nu_k^0 > 0$, have smooth spectral envelopes and different spectral centroids, so as to form a filterbank-like decomposition of the whole spectrum, and $\{e_{km}\}_m$ are the nonnegative coefficients of this decomposition. In this way, it is guaranteed that $w_k(f)$ is a harmonic spectrum of fundamental frequency ν_k^0, with a smooth spectral envelope. If the signal of interest is a music signal, then the order K and the fundamental frequencies ν_k^0 can typically be preset according to the semitone scale; otherwise they have to be estimated along with the other parameters. Two methods were proposed for estimating the coefficients e_{km} and the activations in matrix $\mathbf{H}$ from the observed spectrogram: a space-alternating generalized EM algorithm based on a Gaussian model [Bertin et al., 2010] (cf. Section 2.3.2) and multiplicative update rules (with a faster convergence speed) based either on the IS divergence [Bertin et al., 2010], or more generally on the β-divergence [Vincent et al., 2010].

Hennequin et al. [2010] proposed a similar parameterization of the spectrum vector $\mathbf{w}_k$, considering that har-

monic spectra are formed of a number M of distinct partials:

$$w_k(f) = \sum_m a_k^m \, g_{km}(f), \tag{1.12}$$

where $a_k^m \geq 0$, $g_{km}(f) = g\left(\nu_f - \nu_k^m\right)$, $\nu_f = \frac{f}{F} f_s$ and $\nu_k^m = m\,\nu_k^0$, and $g(\cdot)$ is the spectrum of the analysis window used for computing the spectrogram. Multiplicative update rules based on the β-divergence were proposed for estimating this model. Since this parametric model does not explicitly enforce the smoothness of the spectral envelope, a regularization term promoting this smoothness was added to the β-divergence, resulting in a better decomposition of music spectrograms [Hennequin et al., 2010].

3.1.4 Inharmonicity

When modeling some string musical instruments such as the piano or the guitar, the harmonicity assumption has to be relaxed. Indeed, because of the bending stiffness, the partial frequencies no longer follow an exact harmonic progression, but rather a so-called *inharmonic* progression:

$$\nu_k^m = m\nu_k^0 \sqrt{1 + Bm^2}, \tag{1.13}$$

where m is the partial index, $B > 0$ is the inharmonicity coefficient, and $\nu_k^0 > 0$ is the fundamental frequency of vibration of an ideal flexible string [Rigaud et al., 2013]. Then the spectrum vector $\mathbf{w}_k$ can be parameterized as in (1.12), and all parameters, including the inharmonicity coefficient B, can be estimated by minimizing the β-divergence criterion by means of multiplicative update rules. However, it was observed that the resulting algorithm is very sensitive to initialization (cf. Section 2.3.1). In order to improve the robustness to initialization, the exact parameterization of frequencies ν_k^m in (1.13) was relaxed by considering these frequencies as free parameters, and by adding the following regularization term to the β-divergence criterion: $\sum_{km} |\nu_k^m - m\nu_k^0 \sqrt{1 + Bm^2}|^2$.

3.2 Nonstationarity

In the previous Section 3.1, several methods have been presented for enforcing the harmonicity and the spectral smoothness of vectors $\mathbf{w}_k$ in matrix $\mathbf{W}$, by means of either hard or soft constraints. All these methods assumed that the spectra of the audio events forming the observed spectrogram are stationary. However, many real audio signals are known to be nonstationary: the fundamental frequency, as well as the spectral envelope, may vary over time. In this section, we present some models that aim to represent such nonstationary signals, by allowing the fundamental frequency and spectral envelope parameters to vary over time.

3.2.1 Time-varying fundamental frequencies

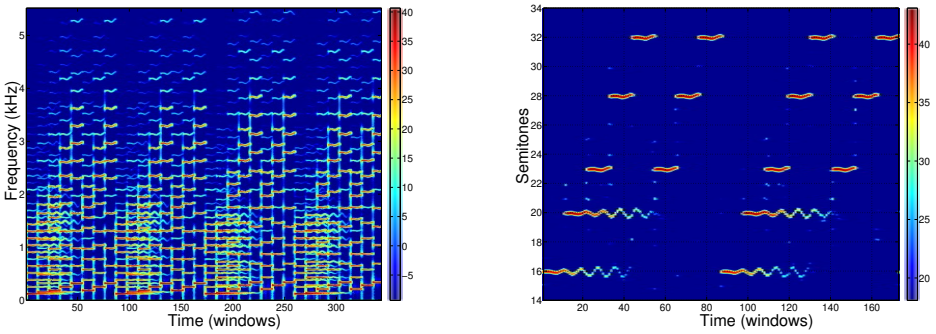
In Section 3.1.3, a harmonic parameterization of vector $\mathbf{w}_k$ was described in (1.12). Hennequin et al. [2010] proposed a straightforward generalization of this model by making the spectral coefficient w_k also depend of time n: $w_k(n, f) = \sum_m a_k^m \, g\left(\nu_f - \nu_k^m(n)\right)$, resulting in a a spectrogram model that is a generalization of NMF: $\widehat{v}(n, f) = \sum_k \widehat{v}_k(n, f)$ with $\widehat{v}_k(n, f) = w_k(n, f)\, h_k(n)$.

Multiplicative update rules based on the β-divergence were proposed for estimating this extended model, along with several regularization terms designed to better fit music spectrograms [Hennequin et al., 2010].

We used this model to decompose the spectrogram of an excerpt (first 4 bars) of the first prelude of Johann Sebastian Bach played by a synthesizer (figure 1.4(a)). A slight vibrato has been added in the played notes to emphasize the variable fundamental frequency estimation. The decomposition uses 72 spectral shapes distributed every semitone. The figure 1.4(b) represents the activations and the fundamental frequencies ν_{kn}^0 obtained. The notes of the prelude appear very clearly, with the vibrato effect.

3.2.2 Time-varying spectral envelopes

Beyond the fundamental frequency, the spectral envelope of freely vibrating harmonic tones (such as those produced by a piano or a guitar) is not constant over time: generally, the upper partials decrease faster than the lower



(a) Original spectrogram

(b) Representation of the activations including the fundamental frequencies (first two measurements of the extract). The color scale is in dB.

Figure 1.4: Decomposition of an excerpt from the first Prelude by Johann Sebastian Bach (figure extracted from Hennequin et al. [2010])

ones. Besides, some sounds such as those produced by a didgeridoo are characterized by a strong resonance in the spectrum that varies over time. Similarly, every time fingerings change on a wind instrument, the shape of the resonating body changes and the resonance pattern is different.

In order to properly model such sounds involving time-varying spectra, Hennequin et al. [2011] proposed to make the activations in vector $\mathbf{h}_k$ not only depend on time, but also on frequency, in order to account for the temporal variations of the spectral envelope of vector $\mathbf{w}_k$. More precisely, the activation coefficient $h_k(n,f)$ is parameterized according to an *Autoregressive Moving Average* (ARMA) model:

$$h_k(n,f) = \sigma_k^2(n) \left| \frac{1 + \sum_{n'} \beta_k(n,n') e^{-2j\pi n' f/F}}{1 - \sum_{n'} \alpha_k(n,n') e^{-2j\pi n' f/F}} \right|^2 \tag{1.14}$$

where $\sigma_k^2(n)$ is the variance parameter at time n, $\alpha_k(n,n')$ denotes the *Autoregressive* (AR) coefficients, and $\beta_k(n,n')$ the *Moving Average* (MA) coefficients, at time n. Then the NMF model is generalized in the following way: $\widehat{v}(n,f) = \sum_k \widehat{v}_k(n,f)$ with $\widehat{v}_k(n,f) = w_k(f)\, h_k(n,f)$. This model is also estimated by minimizing a β-divergence criterion. All parameters, including the ARMA coefficients, are computed by means of multiplicative update rules, without any training. Note that even though the ARMA model of $h_k(n,f)$ in (1.14) is nonnegative, the model coefficients $\alpha_k(n,n')$ and $\beta_k(n,n')$ are not necessarily nonnegative, which means that the multiplicative update rules introduced in Section 2.3.1 were generalized so as to handle these coefficients appropriately [Hennequin et al., 2011].

This algorithm allowed to efficiently represent non-stationary sounds with strong spectral variations, such as the Jew's harp sounds. The Jew's harp is an instrument made of a vibrating metal rod. This rod is placed in the mouth of the instrumentalist who modulates the sound with his mouth. It is thus a harmonic sound (with a fixed fundamental frequency) with a strong resonance varying with time (see the spectrogram in figure 1.5(a)). The decomposition obtained with our algorithm (using a single component) is shown in figures 1.5(b) and 1.5(c) : it shows well the harmonic shape of the spectrum on the one hand and the temporal variations of the resonance on the other hand.

3.2.3 Both types of variations

In order to account for the temporal variations of the fundamental frequency and the spectral envelope jointly, Fuentes et al. [2013] proposed a model called harmonic adaptive latent component analysis. This models falls within the scope of the PLCA framework described in Section 2.2.2: matrix $\widehat{\mathbf{V}}$ is viewed as a discrete probability distribution $P(n,f) = \widehat{v}(n,f)$, and the spectrogram $\mathbf{V}$ is modeled as a histogram: $v(n,f) = \frac{1}{M} \sum_m \delta_{(n_m,f_m)}(n,f)$, where $\{(n_m,f_m)\}_m$ are i.i.d. random vectors distributed according to $P(n,f)$.

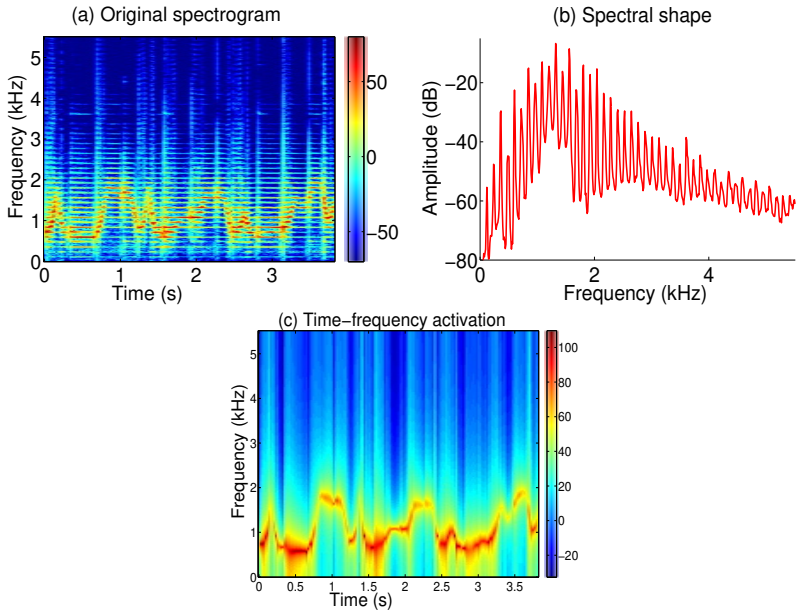


Figure 1.5: Jew's harp sound decomposed with a time-frequency activation parameterized by an ARMA filter of order (1,1) (figure extracted from Hennequin [2010])

In practice, the time-frequency transform used to compute $\mathbf{V}$ is a constant-Q transform. Because this transform involves a log-frequency scale, pitch shifting can be approximated as a translation of the spectrum along this log-frequency axis. Therefore all the notes produced by source j at time n are approximately characterized by a unique template spectrum modeled by a probability distribution $P(\mu \mid j,n)$ (where μ is a frequency parameter), which does not depend on the fundamental frequency. However this distribution depends on time n in order to account for possible temporal variations of the spectral envelope. Besides, the variation of the pitch f_0 of source j over time n is modeled by a probability distribution $P(f_0 \mid j,n)$. Hence the resulting distribution of the shifted frequency $f = \mu + f_0$ is $P(f \mid j,n) = \sum_{f_0} P(f - f_0 \mid j,n)P(f_0 \mid j,n)$. Finally, the presence of source j at time n is characterized by a distribution $P(j,n)$. Therefore the resulting spectrogram corresponds to the probability distribution $P(n,f) = \sum_{jf_0} P(f - f_0 \mid j,n)P(f_0 \mid j,n)P(j,n)$.

In order to enforce both the harmonicity and the smoothness of the spectral envelope, the template spectrum $P(\mu \mid j,n)$ is modeled in the same way as in the first paragraph of Section 3.1.3, as a nonnegative linear combination of K narrowband, harmonic spectral patterns $P(\mu \mid k)$: $P(\mu \mid j,n) = \sum_k P(\mu \mid k)P(k \mid j,n)$, where $P(k \mid j,n)$ is the nonnegative weight of pattern k at time n for source j. Finally, the resulting harmonic adaptive latent component analysis model is expressed as

$$P(n,f) = \sum_{f_0 k j} P(f - f_0 \mid k)P(k \mid j,n)P(f_0 \mid j,n)P(j,n). \tag{1.15}$$

4 Summary

In this chapter, we have shown that NMF is a very powerful model for representing speech and music data. We have presented the mathematical foundations, and described several probabilistic frameworks and various algorithms for computing an NMF. We have also presented some advanced NMF models that are able to more accurately represent audio signals, by enforcing properties such as sparsity, harmonicity and spectral smoothness, and by taking the nonstationarity of the data into account. We have shown that coupled factorizations make it possible to exploit some extra information we may have about the observed signal, such as the musical score. Finally, we have presented several methods that perform dictionary learning for NMF.

The benefits of NMF in comparison with other separation approaches are the capability of performing unsupervised source separation, learning source models from a relatively small amount of material (especially in comparison with *Deep Neural Networks* (DNN)), and easily implementing and adapting the source models and the algorithms. The main downside is the complexity of iterative NMF algorithms. Note that beyond source separation, NMF models have also proved successful in a broad range of audio applications, including automatic music transcription [Smaragdis and Brown, 2003], multipitch estimation [Vincent et al., 2010, Bertin et al., 2010, Fuentes et al., 2013, Benetos et al., 2014], and audio inpainting [Smaragdis et al., 2011].

Bibliography

R. Badeau and A. Drémeau. Variational Bayesian EM algorithm for modeling mixtures of non-stationary signals in the time-frequency domain (HR-NMF). In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pages 6171–6175, May 2013.

R. Badeau, N. Bertin, and E. Vincent. Stability analysis of multiplicative update algorithms and application to non-negative matrix factorization. *IEEE Transactions on Neural Networks*, 21(12):1869–1881, Dec. 2010.

E. Benetos, G. Richard, and R. Badeau. Template adaptation for improving automatic music transcription. In *Proceedings of International Society for Music Information Retrieval Conference*, pages 175–180, Oct. 2014.

N. Bertin. *Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.* PhD thesis, École Nationale Supérieure des Télécommunications, Paris, France, Oct. 2009.

N. Bertin and R. Badeau. Initialization, distances and local minima in audio applications of the non-negative matrix factorization. In *Acoustics'08*, Paris, France, July 2008.

N. Bertin, C. Févotte, and R. Badeau. A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pages 1545–1548, Apr. 2009.

N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 18(3):538–549, Mar. 2010.

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

Z. Chen, A. Cichocki, and T. M. Rutkowski. Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer's disease. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 893–896, Toulouse, France, May 2006.

A. Cichocki, R. Zdunek, and S.-i. Amari. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *Proc. of the 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 32–39, Charleston, Caroline du Sud, USA, Mar. 2006.

A. Cichocki, R. Zdunek, and S.-i. Amari. Nonnegative matrix and tensor factorization. *IEEE Signal Processing Magazine*, 25(1):142–145, Jan. 2008.

A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley, Sept. 2009.

S. Eguchi and Y. Kano. Robustifying maximum likelihood estimation. Technical report, Tokyo Institute of Statistical Mathematics, Tokyo, Japon, 2001. URL `http://www.ism.ac.jp/~eguchi/pdf/Robustify_MLE.pdf`.

J. A. Fessler and A. O. Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing*, 42(10):2664–2677, Oct. 1994.

C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, Sep. 2011.

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

L. Finesso and P. Spreij. Approximate nonnegative matrix factorization via alternating minimization. In *Proceedings of International Symposium on Mathematical Theory of Networks and Systems*, July 2004.

B. Fuentes, R. Badeau, and G. Richard. Harmonic adaptive latent component analysis of audio and application to music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 21(9):1854–1866, Sept. 2013.

R. Hennequin. Rapport à mi-parcours de travaux de thèse. Télécom ParisTech, Apr. 2010.

R. Hennequin, R. Badeau, and B. David. Time-dependent parametric and harmonic templates in non-negative matrix factorization. In *Proceedings of International Conference on Digital Audio Effects*, Sept. 2010.

R. Hennequin, R. Badeau, and B. David. NMF with time-frequency activations to model non-stationary audio events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 19(4):744–753, May 2011.

A. Hurmalainen, R. Saeidi, and T. Virtanen. Similarity induced group sparsity for non-negative matrix factorisation. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pages 4425–4429, Apr. 2015.

P. I. M. Johannesma. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In *Proc. of Symposium on Hearing Theory*, pages 58–69, Eindhoven, Pays-Bas, June 1972.

H. Kameoka, N. Ono, K. Kashino, and S. Sagayama. Complex NMF: A new sparse representation for acoustic signals. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pages 3437–3440, April 2009.

Y.-D. Kim and S. Choi. A method of initialization for nonnegative matrix factorization. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 537–540, Honolulu, Hawaii, USA, Apr. 2007.

R. Kompass. A generalized divergence measure for nonnegative matrix factorization. *Neural Computation*, 19(3): 780–791, Mar. 2007.

J. P. Lasalle. *The stability and control of discrete processes*. Springer-Verlag, New York, NY, USA, 1986.

H. Laurberg, M. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen. Theorems on positive data: On the uniqueness of NMF. *Computational Intelligence and Neuroscience*, 2008, 2008. Article ID 764206, 9 pages.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755): 788–791, Oct. 1999.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proceedings of Neural Information Processing Systems*, pages 556–562, Dec. 2001.

A. Lefevre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pages 21–24, May 2011.

C.-J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19(10):2756–2779, Oct. 2007.

A. Liutkus and R. Badeau. Generalized Wiener filtering with fractional power spectrograms. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pages 266–270, Apr. 2015.

A. Liutkus, D. Fitzgerald, and R. Badeau. Cauchy nonnegative matrix factorization. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2015.

B. C. Moore and B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America (JASA)*, 74(3):750–753, Sept. 1983.

M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama. Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence. In *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, pages 283–288, 2010.

J. Nikunen and T. Virtanen. Noise-to-mask ratio minimization by weighted non-negative matrix factorization. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 25–28, Dallas, Texas, USA, Mar. 2010.

S. A. Raczyński, N. Ono, and S. Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proc. of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, Vienne, Autriche, Sept. 2007.

F. Rigaud, B. David, and L. Daudet. A parametric model and estimation techniques for the inharmonicity and tuning of the piano. *Journal of the Acoustical Society of America*, 133(5):3107–3118, May 2013.

M. N. Schmidt and H. Laurberg. Non-negative matrix factorization with Gaussian process priors. *Computational Intelligence and Neuroscience*, 2008:1–10, 2008. Article ID 361705.

M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008:1–8, 2008. Article ID 947438.

U. Simsekli and A. T. Cemgil. Markov chain Monte Carlo inference for probabilistic latent tensor factorization. In *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2012.

R. Singh, B. Raj, and P. Smaragdis. Latent-variable decomposition based dereverberation of monaural andmulti-channel signals. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1914–1917, Dallas, Texas, USA, Mar. 2010.

P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Proc. of International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 494–499, Grenade, Espagne, Sept. 2004.

P. Smaragdis. Relative pitch tracking of multiple arbitrary sounds. *Journal of the Acoustical Society of America (JASA)*, 125(5):3406–3413, May 2009.

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, Oct. 2003.

P. Smaragdis, B. Raj, and M. Shashanka. Missing data imputation for time-frequency representations of audio signals. *Journal of Signal Processing Systems*, 65:361–370, Aug 2011.

E. Vincent, N. Bertin, and R. Badeau. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 109–112, Las Vegas, Nevada, USA, Apr. 2008.

E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, Mar. 2010.

T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, Mar. 2007.

T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pages 1825–1828, Apr. 2008.

S. Wild, J. Curry, and A. Dougherty. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, 37(11):2217–2232, Nov. 2004.

Y. Zhang and Y. Fang. A NMF algorithm for blind separation of uncorrelated signals. In *Proc. of IEEE International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, pages 999–1003, Beijing, Chine, Nov. 2007.