Single-channel audio source separation with NMF: divergences, constraints and algorithms

Cédric Févotte, Emmanuel Vincent, Alexey Ozerov

Chapter 1
Single-channel audio source separation with NMF: divergences, constraints and algorithms

Cédric Févotte[1], Emmanuel Vincent[2], and Alexey Ozerov[3]

Abstract

Spectral decomposition by nonnegative matrix factorisation (NMF) has become state-of-the-art practice in many audio signal processing tasks, such as source separation, enhancement or transcription. This chapter reviews the fundamentals of NMF-based audio decomposition, in unsupervised and informed settings. We formulate NMF as an optimisation problem and discuss the choice of the measure of fit. We present the standard majorisation-minimisation strategy to address optimisation for NMF with common β-divergence, a family of measures of fit that takes the quadratic cost, the generalised Kullback-Leibler divergence and the Itakura-Saito divergence as special cases. We discuss the reconstruction of time-domain components from the spectral factorisation and present common variants of NMF-based spectral decomposition: supervised and informed settings, regularised versions, temporal models.

1.1 Introduction

Data is often available in matrix form $\mathbf{V}$, where columns $\mathbf{v}_n$ are data samples and rows are features. Processing such data often entails finding a factorisation of the matrix $\mathbf{V}$ into two unknown matrices $\mathbf{W}$ and $\mathbf{H}$ such that

$$\mathbf{V} \approx \hat{\mathbf{V}} \stackrel{\text{def}}{=} \mathbf{WH}. \tag{1.1}$$

In the approximation (1.1), $\mathbf{W}$ acts a dictionary of recurring patterns, which is characteristic of the data, and every column $\mathbf{h}_n$ of $\mathbf{H}$ contains the *decomposition* or *ac-*

[1] CNRS & IRIT, Toulouse, France
[2] Inria, 54600 Villers-lès-Nancy, France
[3] Technicolor, Rennes, France

tivation coefficients that approximate every $\mathbf{v}_n$ onto the dictionary. In the following we will refer to $\mathbf{W}$ as the *dictionary* and to $\mathbf{H}$ as the *activation matrix*. The data matrix $\mathbf{V}$ is of dimensions $F \times N$ and the common dimension of $\mathbf{W}$ and $\mathbf{H}$ is denoted K, often referred to as the rank of the factorisation (which might differ from the actual mathematical rank of $\mathbf{V}$).

In the literature, the problem of obtaining the factorisation (1.1) can appear under other domain-specific names such as *dictionary learning*, *low-rank approximation*, *factor analysis* or *latent semantic analysis*. Many forms of factorisation (1.1) have been considered. The most notorious and ancient one is Principal Component Analysis (PCA) [1] which simply minimises the quadratic cost between $\mathbf{V}$ and its approximate $\mathbf{WH}$, where all matrices are treated as real-valued. Independent Component Analysis (ICA) [2] is a major variant of PCA in which the rows of $\mathbf{H}$ are constrained to be mutually independent. Sparse coding [3] and many recent dictionary learning [4] approaches impose some form of sparsity of the activation matrix. , the main topic of this chapter, is dedicated to nonnegative data and imposes nonnegativity of the factors $\mathbf{W}$ and $\mathbf{H}$.

Early work on NMF has appeared in applied algebra (under various names) and more notably in chemometrics [5], but it fully came to maturation with the seminal paper of Lee and Seung, published in *Nature* in 1999 [6]. Like PCA, NMF consists of minimising an error of fit between $\mathbf{V}$ and its approximate $\mathbf{WH}$, but subject to nonnegativity of the values of $\mathbf{W}$ and $\mathbf{H}$. The nonnegativity of $\mathbf{W}$ ensures the *interpretability* of dictionary, in the sense that the extracted patterns $\mathbf{w}_k$ (the columns of $\mathbf{W}$) remain nonnegative, like the data samples. The nonnegativity of $\mathbf{H}$ ensures that $\mathbf{WH}$ is nonnegative, like $\mathbf{V}$, but is also shown to induce a *part-based representation*, in stark contrast with plain PCA that leads to more global or *holistic* representations (where every pattern attempts to generalise as much as possible the whole dataset). Because subtractive combinations are forbidden, the approximate $\mathbf{Wh}_n$ to every sample $\mathbf{v}_n$ can only be formed from building blocks, and thus the estimated patterns tend to be parts of data.

Following the work of Lee and Seung, NMF became an increasingly popular data analysis tool and has been used in many fields. In particular, it has led to important breakthroughs in text retrieval (based on the decomposition of a *bag-of-words* representation [7]), collaborative filtering (completion of missing ratings in users $\times$ items matrices [8]) or spectral unmixing. In the latter case, NMF is for example used in chemical spectroscopy [5], remote sensing (for unmixing of hyperspectral electromagnetic data) [9] and most notably audio signal processing [10]. The seminal work of Smaragdis *et al.* [10] has initiated an important thread of NMF-based contributions in music transcription, source separation, speech enhancement, etc. The common principle of all these works is the nonnegative decomposition of the spectrogram of the observed signal onto a dictionary of elementary spectral components, representative of building sound units (notes, chords, percussive sounds, or more complex adaptive structures). This general architecture is detailed in Section 1.2. It describes in particular popular NMF models and means of obtaining the factorisation, by optimisation of a cost function. Then it describes how to reconstruct elementary sound components from the nonnegative factorisation of the

spectrogram. This blind decomposition might fail to return adequate and useful results when dealing with complex multi-source signals and the system needs to be "guided" with prior information. Such advanced decompositions for source separation will be covered in Section 1.3. Section 1.4 concludes.

1.2 Signal decomposition by NMF

The general principle of NMF-based audio spectral analysis is depicted in Fig. 1.1. It shows how NMF has the capability of unmixing superimposed spectral components. This is in contrast for example with the Gaussian Mixture Model (GMM), a clustering model that is not designed to handle composite data. In the GMM, each data sample can only be in one among several states. As such, the occurrence of mixed frames in the data represented in Fig. 1.1 (3rd to 5th samples) would count as one state, along the two other states corresponding to pure spectra (red and green). The nonnegativity of $\mathbf{H}$ encourages so-called *part-based* representations. Because subtractive combinations of dictionary elements are forbidden, the dictionary $\mathbf{W}$ tends to contain elementary building units. This is a welcome property for analysis tasks such as music transcription or source separation. In contrast, a method such as PCA would instead produce an orthogonal dictionary with a more *holistic* value, that compresses more efficiently the entire dataset. The difference between PCA, NMF and vector quantisation is remarkably illustrated in [6] with comparative experiments using a set of face images. It is shown that where PCA returns *eigenfaces* (sort of template faces), NMF can efficiently capture parts of faces (noise, eyes, etc.). Figure 1.2 displays the result of NMF applied to the spectrogram of a short piano sequence; see [10, 11] for further illustration on small-scale examples.

1.2.1 NMF by optimisation

The factorisation (1.1) is usually sought after through the minimisation problem

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{WH}) \text{ subject to } \mathbf{W} \geq 0, \mathbf{H} \geq 0 \tag{1.2}$$

where the notation $\mathbf{A} \geq 0$ expresses nonnegativity of the entries of matrix $\mathbf{A}$ (and not semidefinite positiveness), and where $D(\mathbf{V}|\mathbf{WH})$ is a separable such that

$$D(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}) \tag{1.3}$$

where $d(x|y)$ is a scalar cost function. What we intend by "" is a positive function of $y \in \mathbb{R}_+$ given $x \in \mathbb{R}_+$, with a single minimum for $x = y$.

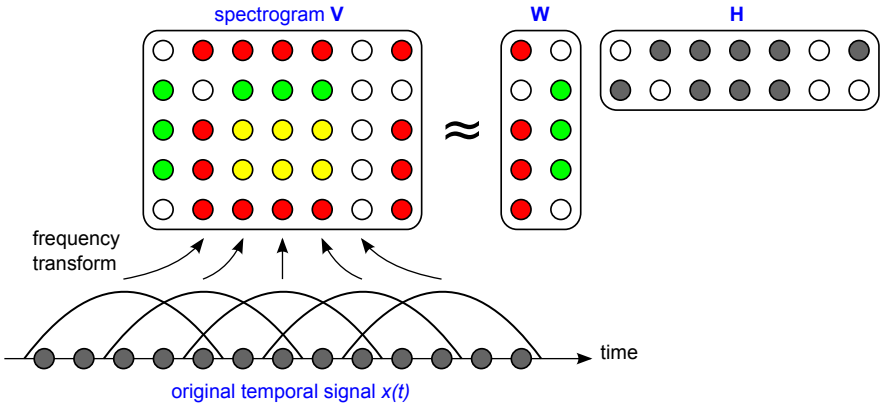

Fig. 1.1 NMF-based audio spectral analysis. A short-time frequency transform, such as the magnitude or power short-time Fourier transform, is applied to the original time-domain signal $x(t)$. The resulting nonnegative matrix is factorised into the nonnegative matrices **W** and **H**. In this schematic example, the red and green elementary spectra are unmixed and extracted into the dictionary matrix **W**. The activation matrix **H** returns the mixing proportions of each time-frame (a column of **W**).

The quadratic cost function $d_Q(x|y) = \frac{1}{2}(x-y)^2$ is a popular choice when dealing with real numbers. It underlies an additive Gaussian noise model and enjoys convenient mathematical properties for estimation and optimisation problems. For that same reason, it is a less natural choice for nonnegative data because it may generate negative values. Many other choices have been considered in the NMF literature, in particular under the influence of Cichocki et al. Two popular families of NMF cost functions are the α-divergence [12] and the β-divergence [13, 14, 15], themselves connected to the wider families of Csiszár or Bregman divergences, see, e.g., [13] and [16] in the context of NMF. The β-divergence in particular has enjoyed a certain success in audio signal processing. It can be defined as [17, 18]

$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}\right), & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\log\frac{x}{y} - x + y = d_{KL}(x|y), & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 = d_{IS}(x|y), & \beta = 0 \end{cases} \tag{1.4}$$

The limit cases $\beta = 0$ and $\beta = 1$ correspond to the Itakura-Saito (IS) and generalised Kullback-Leibler (KL) divergences, respectively. The case $\beta = 2$ corresponds to the quadratic cost $d_Q(x|y)$. The β-divergence forms a continuous family of cost functions that smoothly interpolates between the latter three well-known cases. As noted in [11, 15], a noteworthy property of the β-divergence is its behaviour w.r.t. the scale of the data, as the following equation holds for any value of β:

$$d_\beta(\lambda x|\lambda y) = \lambda^\beta \, d_\beta(x|y). \tag{1.5}$$

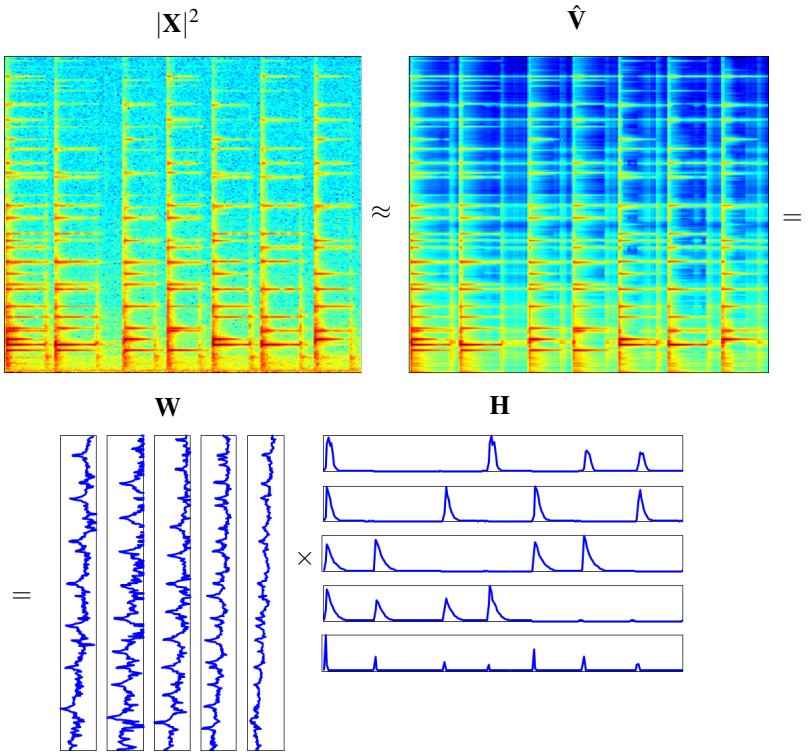

Fig. 1.2 NMF applied to the spectrogram of a short piano sequence composed of four notes. (Data used from [11]).

As noted in [11], this implies that factorisations obtained with $\beta > 0$ (such as with the quadratic cost or the KL divergence) will rely more heavily on large data values and less precision is to be expected in the estimation of the low-power components, and conversely factorisations obtained with $\beta < 0$ will rely more heavily on small data values. The IS divergence ($\beta = 0$) is scale-invariant, i.e., $d_{IS}(\lambda x|\lambda y) = d_{IS}(x|y)$, and is the only one in the family of β-divergences to possess this property. Factorisations with small positive values of β are relevant to decomposition of audio spectra, which typically exhibit exponential power decrease along frequency f and also usually comprise low-power transient components such as note attacks together with higher power components such as tonal parts of sustained notes. For example, [11] presents the results of the decomposition of a piano power spectrogram with IS-NMF and shows that components corresponding to very low residual noise and hammer hits on the strings are extracted with great accuracy, while these components are either ignored or severely degraded when using Euclidean or KL divergences. Similarly, the value $\beta = 0.5$ is advocated by [19, 20]

and has been shown to give optimal results in music transcription based on NMF of the magnitude spectrogram by [21].

1.2.2 Composite models

NMF with the β-divergence as formulated in the previous section fails to give a probabilistic understanding of the modelling assumptions. As a matter of fact, the β-divergence acts as a pseudo-likelihood for the so-called Tweedie distribution, a member of the exponential family, parametrised with respect to its mean, i.e., such that [22]

$$\mathrm{E}\left[\mathbf{V}|\mathbf{WH}\right] = \mathbf{WH}. \tag{1.6}$$

In particular, the values $\beta = 0, 1, 2$ underlie multiplicative Gamma observation noise ($v_{fn} = [\mathbf{WH}]_{fn}.\varepsilon_{fn}$), Poisson noise ($v_{fn} \sim Po([\mathbf{WH}]_n)$) and Gaussian additive observation noise ($v_{fn} = [\mathbf{WH}]_{fn} + \varepsilon_{fn}$), respectively (see the Appendix for the definitions of the distributions involved).

These probabilistic models characterise the magnitude or power spectrogram $\mathbf{V}$ but do not explicitly characterise the composite structure of sound that is generally looked after in NMF-based decomposition. As such, the *Gaussian Composite Model* (GCM) was introduced in [11] to remedy this limitation. Denoting by x_{fn} the complex-valued coefficients of the short-time Fourier transform (STFT), the GCM is defined by

$$x_{fn} = \sum_k c_{k,fn}, \tag{1.7}$$

$$c_{k,fn} \sim N_c(0, w_{fk}h_{kn}), \tag{1.8}$$

where $N_c(\mu, \lambda)$ refers to the circular complex-valued normal distribution defined in the Appendix. The composite structure of sound (i.e., the superimposition of elementary components) is made explicit by (1.7). Then, (1.8) states that the k^{th} elementary component $c_{k,fn}$ is the expression of the k^{th} the spectral template $\mathbf{w}_k$ amplitude-modulated in time by the activation coefficient h_{kn}. The latent components may also be marginalised from the model to yield more simply

$$x_{fn} \sim N_c(0, [\mathbf{WH}]_{fn}). \tag{1.9}$$

With the uniform phase assumption that defines the circular complex-valued normal distribution, (1.9) itself reduces to

$$v_{fn} = [\mathbf{WH}].\varepsilon_{fn}, \tag{1.10}$$

where $v_{fn} = |x_{fn}|^2$ (the power spectrogram) and ε_{fn} has an exponential distribution with expectation 1 (i.e., using the notations defined in the Appendix, $\varepsilon_{fn} \sim G(1,1)$).

As such, the GCM is tightly connected to the multiplicative Gamma noise model, and we may easily find that

$$-\log p(\mathbf{X}|\mathbf{WH}) = D_{IS}(|\mathbf{X}|^2|\mathbf{WH}) + cst. \tag{1.11}$$

(1.11) shows that factorising the power spectrogram $\mathbf{V} = |\mathbf{X}|^2$ with the IS divergence is equivalent to performing maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in the GCM model defined by (1.7) and (1.8). Given estimates of $\mathbf{W}$ and $\mathbf{H}$ (using for example the algorithm presented in the following section), reconstruction of the latent components $c_{k,fn}$ can be done with any estimator. For example, the Minimum Mean Squares Error (MMSE) estimator is given by the so-called Wiener filter

$$\hat{c}_{k,fn} = \mathrm{E}\left[c_{k,fn}|\mathbf{W},\mathbf{H}\right] = \frac{w_{fk}h_{kn}}{[\mathbf{WH}]_{fn}}x_{fn} \tag{1.12}$$

By construction, the component estimates satisfy $x_{fn} = \sum_k \hat{c}_{k,fn}$. The estimated component STFTs $\hat{\mathbf{C}}_k = \{c_{k,fn}\}_{fn}$ can then be inverse-transformed (using a standard overlap-add procedure) to yield time-domain estimates $\hat{c}_k(t)$ such that $x(t) = \sum_k \hat{c}_k(t)$.

Besides the GCM, other composite interpretations of known NMF models have been proposed in the literature [23]. For example, the Poisson-NMF model

$$v_{fn} \sim Po([\mathbf{WH}]_{fn}) \tag{1.13}$$

is equivalent to

$$x_{fn} = \sum_k c_{k,fn}, \tag{1.14}$$

$$c_{k,fn} \sim Po(w_{fk}h_{kn}). \tag{1.15}$$

It turns out the MMSE estimator of the latent components is again given by (1.12). It is easily shown that

$$-\log p(\mathbf{V}|\mathbf{WH}) = D_{KL}(\mathbf{V}|\mathbf{WH}) + cst \tag{1.16}$$

so that maximum-likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in model (1.17) is equivalent to NMF with the generalised KL divergence [24, 25, 11]. A closely related model is PLSA [7] / PLCA [26] which writes

$$\mathbf{v}_n \sim M(\sum_f v_{fn}, \mathbf{W}\mathbf{h}_n), \tag{1.17}$$

where $M(L,\mathbf{p})$ refers to the multinomial distribution defined in the Appendix and the columns of $\mathbf{W}$ and $\mathbf{H}$ are constrained to sum to 1. PLSA/PLCA can also be shown to be equivalent to a generative model that involves multinomial latent components. PLCA is equivalent to NMF with a weighted KL divergence, such that

$$-\log p(\mathbf{V}|\mathbf{WH}) = \sum_n \|\mathbf{v}_n\|_1 D_{KL}\left(\frac{\mathbf{v}_n}{\|\mathbf{v}_n\|_1}|\mathbf{W}\mathbf{h}_n\right). \quad (1.18)$$

Poisson-NMF and PLCA are also popular models for audio spectrogram decomposition. This is because the KL divergence (used with the magnitude spectrogram $\mathbf{V} = |\mathbf{X}|$) has been experimentally proven to be also a reasonable measure of fit for audio spectral factorisation [27, 28]. However, from a probabilistic generative point of view, the Poisson-NMF and PLCA models are unreasonable because they generate integer values that do not comply with the real-valued nature of spectrograms (as a matter of fact, Poisson-NMF and PLSA/PLCA have been originally designed for count data [7, 24]).

1.2.3 Majorisation-minimisation

The very large majority of NMF algorithms resort to block-coordinate descent to address problem (1.2). This means the variables $\mathbf{W}$ and $\mathbf{H}$ are updated in turn until a stationary point of $C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{WH})$ is reached. Because $C(\mathbf{W},\mathbf{H})$ is jointly non-convex in $\mathbf{W}$ and $\mathbf{H}$, the stationary point may be not a global minimum (and possibly not even a local minimum). As such, initialisation is an important issue in NMF and running the algorithm from different starting points is usually advised. It is also easy to see that the updates of $\mathbf{W}$ and $\mathbf{H}$ are essentially the same by transposition ($\mathbf{V} \approx \mathbf{WH} \Leftrightarrow \mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T$). As such we may restrict our study to the update of $\mathbf{H}$ given $\mathbf{W}$:

$$\min_{\mathbf{W},\mathbf{H}} C(\mathbf{H}) \stackrel{\text{def}}{=} D(\mathbf{V}|\mathbf{WH}) \text{ subject to } \mathbf{H} \geq 0 \quad (1.19)$$

For the divergences considered in Section 1.2.1, a standard approach to the conditional updates of $\mathbf{W}$ and $\mathbf{H}$ is . Generally speaking, MM consists in optimising iteratively an easier-to-minimise tight upper bound of the original objective function $C(\mathbf{H})$ [29].

Denote by $\tilde{\mathbf{H}}$ the estimate of $\mathbf{H}$ at current iteration. The first step of MM consists in building an upper bound $G(\mathbf{H}|\tilde{\mathbf{H}})$ of $C(\mathbf{H})$ which is tight for $\mathbf{H} = \tilde{\mathbf{H}}$, i.e., $C(\mathbf{H}) \leq G(\mathbf{H}|\tilde{\mathbf{H}})$ for all $\mathbf{H}$ and $C(\tilde{\mathbf{H}}) = G(\tilde{\mathbf{H}}|\tilde{\mathbf{H}})$. The second step consists in minimising the bound w.r.t. $\mathbf{H}$, producing a valid descent algorithm. Indeed, at iteration $i+1$, it holds by construction that $C(\mathbf{H}^{(i+1)}) \leq G(\mathbf{H}^{(i+1)}|\mathbf{H}^{(i)}) \leq G(\mathbf{H}^{(i)}|\mathbf{H}^{(i)}) = C(\mathbf{H}^{(i)})$. The bound $G(\mathbf{H}|\tilde{\mathbf{H}})$ is often referred to as *auxiliary function*. The principle of MM is illustrated in Fig. 1.3.

The question now boils down to whether the construction of such an upper bound, which is amenable to optimisation, is possible. Fortunately, the answer is yes for many divergences, and in particular for the β-divergence discussed in Section 1.2.1. The trick is to decompose $C(\mathbf{H})$ into the sum of a convex part and a concave part and to upper-bound each part separately (the concave part is actually inexistent for $1 \leq \beta \leq 2$ where the β-divergence is convex w.r.t. its second argument). The con-

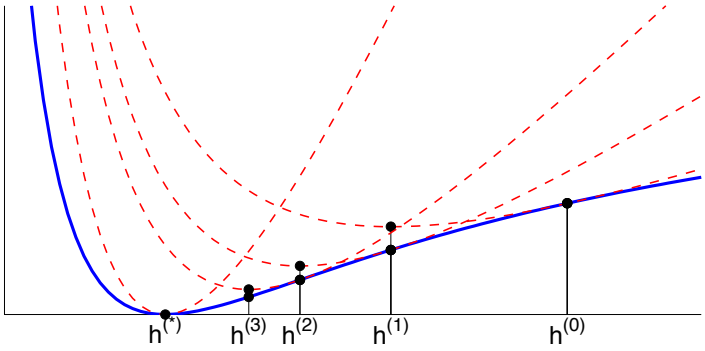

Fig. 1.3 An illustration of the MM principle on a unidimensional problem. Given a current estimate of $\mathbf{W}$, the blue curve acts as the objective function $C(\mathbf{H}) = D(\mathbf{V}|\mathbf{WH})$ to be minimised with respect to $\mathbf{H}$. The MM approach relies on the iterative minimisation of tight upper bounds (dashed red curves). The algorithm is initialised at $\mathbf{H}^{(0)}$, at which the first upper bound is minimised during the first iteration to yield $\mathbf{H}^{(1)}$, and so on until convergence. (Reproduced from [30])

vex part is majorised using Jensen's inequality (the definition of convexity) and the concave part is majorised using the tangent inequality. The two separate bounds are summed and the resulting (convex) auxiliary function turns out to have a closed-form minimiser. For illustration, we address the case of NMF with the Itakura-Saito divergence. The more general β-divergence case is addressed in details in [15].

A special case: NMF with the Itakura-Saito divergence

Choosing the IS divergence as the measure of fit and addressing the update of $\mathbf{H}$, our goal is to minimise the objective function given by

$$C(\mathbf{H}) = \sum_{fn} \left(\frac{v_{fn}}{[\mathbf{WH}]_{fn}} - \log \frac{v_{fn}}{[\mathbf{WH}]_{fn}} - 1 \right) \tag{1.20}$$

$$= \sum_{fn} \left(\frac{v_{fn}}{[\mathbf{WH}]_{fn}} + \log [\mathbf{WH}]_{fn} \right) + cst \tag{1.21}$$

where cst is a term which is constant w.r.t. $\mathbf{H}$. As such, $C(\mathbf{H})$ can be written as the sum of a convex term $\overset{\smile}{C}(\mathbf{H}) = \sum_{fn} \frac{v_{fn}}{[\mathbf{WH}]_{fn}}$ and a concave term $\overset{\frown}{C}(\mathbf{H}) = \sum_{fn} \log [\mathbf{WH}]_{fn}$. By convexity of $f(x) = 1/x$ for $x \geq 0$ and Jensen's inequality it holds that

$$f(\sum_k \lambda_k x_k) \leq \sum_k \lambda_k f(x_k) \tag{1.22}$$

for any $x_k, \lambda_k \geq 0$ such that $\sum_k \lambda_k = 1$. As such, it holds that

$$\breve{C}(\mathbf{H}) = \sum_{fn} \frac{v_{fn}}{\sum_k \frac{w_{fkh_{kn}}}{\lambda_{fkn}} \lambda_{fkn}} \leq \sum_{fn} v_{fn} \sum_k \frac{\lambda_{fkn}^2}{w_{fk} h_{kn}}, \tag{1.23}$$

for any $\lambda_{fkn} \geq 0$ such that $\sum_k \lambda_{fkn} = 1$. Choosing

$$\lambda_{fkn} = \frac{w_{fk}\tilde{h}_{kn}}{[\mathbf{W}\tilde{\mathbf{H}}]_{fn}} \tag{1.24}$$

and denoting by $\breve{G}(\mathbf{H}|\tilde{\mathbf{H}})$ the right-hand side of (1.23), it can be easily checked that $\breve{G}(\mathbf{H}|\tilde{\mathbf{H}})$ is an auxiliary function for $\breve{C}(\mathbf{H})$.

Now, by concavity of $\widehat{C}(\mathbf{H})$ and the tangent inequality applied at $\mathbf{H} = \tilde{\mathbf{H}}$, we may write

$$\widehat{C}(\mathbf{H}) \leq \widehat{C}(\tilde{\mathbf{H}}) + \sum_{kn} [\nabla \widehat{C}(\tilde{\mathbf{H}})]_{kn} (h_{kn} - \tilde{h}_{kn}) \tag{1.25}$$

Using the chain rule, the gradient term is found to be

$$[\nabla \widehat{C}(\tilde{\mathbf{H}})]_{kn} = \sum_f \frac{w_{fk}}{[\mathbf{W}\tilde{\mathbf{H}}]_{fn}}. \tag{1.26}$$

By construction, the right hand side of (1.25) defines an auxiliary function $\widehat{G}(\mathbf{H}|\tilde{\mathbf{H}})$ of $\widehat{C}(\mathbf{H})$. Assembling $\breve{G}(\mathbf{H}|\tilde{\mathbf{H}})$ and $\widehat{G}(\mathbf{H}|\tilde{\mathbf{H}})$ defines an auxiliary function $G(\mathbf{H}|\tilde{\mathbf{H}})$ of $C(\mathbf{H})$. The auxiliary function $G(\mathbf{H}|\tilde{\mathbf{H}})$ is convex by construction. Computing and cancelling its gradient leads to

$$h_{kn} = \tilde{h}_{kn} \left(\frac{\sum_f w_{fk} v_{fn} [\mathbf{W}\tilde{\mathbf{H}}]^{-2}}{\sum_f w_{fk} [\mathbf{W}\tilde{\mathbf{H}}]^{-1}}. \right)^{\frac{1}{2}} \tag{1.27}$$

Because the new update is found by multiplying the previous update with a correcting factor, the induced algorithm is coined "multiplicative". Because the correcting factor is nonnegative, nonnegativity of the updates is ensured along the iterations, given positive initialisations. [15] proves that dropping the exponent $\frac{1}{2}$ in (1.27) produces an accelerated descent algorithm. The update (1.27) can then be written in algorithmic form using matrix operations as

$$\mathbf{H} \leftarrow \mathbf{H} \circ \frac{\mathbf{W}^T \left((\mathbf{W}\mathbf{H})^{\circ[-2]} \circ \mathbf{V} \right)}{\mathbf{W}^T (\mathbf{W}\mathbf{H})^{\circ[-1]}} \tag{1.28}$$

where the notation $\circ$ denotes MATLAB-like entry-wise multiplication/exponentiation and the fraction bar denotes entry-wise division. By exchangeability of $\mathbf{W}$ and $\mathbf{H}$ by transposition, the update rule for $\mathbf{W}$ is simply given by

$$\mathbf{W} \leftarrow \mathbf{W} \circ \frac{\left((\mathbf{W}\mathbf{H})^{\circ[-2]} \circ \mathbf{V} \right) \mathbf{H}^T}{(\mathbf{W}\mathbf{H})^{\circ[-1]} \mathbf{H}^T} \tag{1.29}$$

The two updates (1.28) and (1.29) are applied in turn until a convergence criterion is met. The two updates have linear complexity per iteration, are free of tuning parameters and are very easily implemented.

As detailed in [15], these derivations can easily be extended to the more general case of NMF with the β-divergence. The resulting updates generalise (1.28) and (1.29) and can be written as

$$\mathbf{H} \leftarrow \mathbf{H} \circ \frac{\mathbf{W}^T \left((\mathbf{W}\mathbf{H})^{\circ[\beta-2]} \circ \mathbf{V} \right)}{\mathbf{W}^T (\mathbf{W}\mathbf{H})^{\circ[\beta-1]}} \tag{1.30}$$

$$\mathbf{W} \leftarrow \mathbf{W} \circ \frac{\left((\mathbf{W}\mathbf{H})^{\circ[\beta-2]} \circ \mathbf{V} \right) \mathbf{H}^T}{(\mathbf{W}\mathbf{H})^{\circ[\beta-1]} \mathbf{H}^T} \tag{1.31}$$

1.3 Advanced decompositions for source separation

In the previous sections, we described the elementary principles of signal decomposition by NMF. The direct application of these principles leads to so-called *unsupervised NMF*, where both the dictionary and the activation coefficients are estimated from the signal to be separated. This approach yields interesting and useful results on toy data. For real audio signals, however, each sound source rarely consists of a single NMF component. For instance, a music source typically involves several notes with different pitches, while a speech source involves several phonemes. Various techniques have been proposed to classify or to cluster individual NMF components into sources [31, 32]. Nevertheless, several issues remain: the learned components may overfit the test signal, several sources may share similar dictionary elements, and the elegance of NMF is lost. These issues have called for more advanced treatments incorporating prior information about the properties of audio sources in general and/or in a specific signal [33].

1.3.1 Pre-specified dictionaries

1.3.1.1 Supervised NMF

So-called is the simplest such treatment. It assumes that each source is characterised by a fixed source-specific dictionary and only the activation coefficients must be estimated from the signal to be separated [34]. Let us assume that the sources are indexed by $j \in \{1, \ldots, J\}$ and denote by $\mathbf{W}_j$ and $\mathbf{H}_j$ the dictionary and the activation matrix associated with source j. The mixture spectrogram $\mathbf{V}$ can then be expressed as in (1.1) where

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 & \cdots & \mathbf{W}_J \end{pmatrix} \tag{1.32}$$

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_J \end{pmatrix} \tag{1.33}$$

result from the concatenation of the source-specific dictionaries and activation matrices. Given the dictionaries $\mathbf{W}_1, \ldots, \mathbf{W}_J$ of all sources, the activation matrices $\mathbf{H}_1, \ldots, \mathbf{H}_J$ can be estimated by applying, for instance, the optimisation procedure described in Section 1.2. The standard multiplicative update with the β-divergence can be equivalently rewritten in terms of each $\mathbf{H}_j$ as

$$\mathbf{H}_j \leftarrow \mathbf{H}_j \circ \frac{\mathbf{W}_j^T((\mathbf{WH})^{\circ[\beta-2]} \circ \mathbf{V})}{\mathbf{W}_j^T(\mathbf{WH})^{\circ[\beta-1]}}. \tag{1.34}$$

Note that, because $\mathbf{W}$ is here fixed, in the case when the cost function is strictly convex ($1 \leq \beta \leq 2$), the resulting update is guaranteed to converge to a global minimum. Eventually, the complex-valued spectrogram $\mathbf{S}_j$ of each source can be estimated by Wiener filtering as

$$\mathbf{S}_j = \frac{\mathbf{W}_j\mathbf{H}_j}{\mathbf{WH}} \circ \mathbf{X}. \tag{1.35}$$

This is equivalent to extracting the signal corresponding to all NMF components in Section 1.2.2 and summing the extracted signals associated with each source. A variant of supervised NMF called assumes that a pre-specified dictionary is available for a subset of sources only and that the remaining sources are jointly represented by an additional dictionary which is estimated from the signal to be separated together with the activation matrices of all sources [35].

In order to apply supervised or semi-supervised NMF, one must design source-specific dictionaries in the first place. This is achieved by learning each dictionary from isolated sounds (e.g., individual notes) or continuous recordings from the desired source. The amount of training data is typically assumed to be large, so that large dictionaries containing hundreds or thousands of components can be trained. Three families of *nonnegative dictionary learning* methods can be found in the literature, which operate by applying NMF or selecting exemplars from the training signals, respectively.

Early dictionary learning methods were based on applying NMF to the training signals [34, 36]. Denoting by $\mathbf{V}_j$ the spectrogram resulting from the concatenation of all training signals for source j, this data can be factorised as

$$\mathbf{V}_j \approx \mathbf{W}_j\mathbf{H}_j. \tag{1.36}$$

The activation matrix $\mathbf{H}_j$ is discarded, while the dictionary $\mathbf{W}_j$ is kept and used together with the dictionaries for the other sources for separation. This method suffers from one major limitation: unless regularisation such as sparsity is enforced (see

Section 1.3.2), the number of dictionary elements must be smaller than the number of frequency bins. As a consequence, each dictionary element encodes widely different source spectra and it may not account well for the source characteristics. For instance, it has been shown that small dictionaries tend to represent the spectral envelope of the sources but to discard pitch characteristics, which are essential for separation. In order to address this issue, it was recently proposed to construct the dictionary from exemplars, i.e., spectra (columns) selected from the full training set $\mathbf{V}_j$. The number of dictionary elements then becomes unlimited and each element represents a single spectrum at a time, so that all characteristics of the desired source are preserved. If the training set is not too large, $\mathbf{W}_j = \mathbf{V}_j$ itself might be used as the dictionary [37]. Alternatively, the dictionary may be constructed by selecting [38] or clustering [39] the columns of $\mathbf{V}_j$. The selection can be random or exploit prior information about, e.g., the phoneme or the note corresponding to each frame.

1.3.1.2 Convolutive NMF

In [36, 40, 41], the concept of nonnegative dictionary learning was extended to spectrogram patches. The original NMF model in (1.1) can be rewritten in each time frame n as

$$\mathbf{v}_n \approx \mathbf{W}\mathbf{h}_n = \sum_{k=1}^{K} \mathbf{w}_k h_{kn}. \tag{1.37}$$

After replacing each single-frame spectrum $\mathbf{w}_k$ by a spectrogram patch consisting of L consecutive frames

$$\mathbf{W}_k = \begin{pmatrix} \mathbf{w}_{k,0} & \cdots & \mathbf{w}_{k,L-1} \end{pmatrix}, \tag{1.38}$$

this model can be extended into

$$\mathbf{v}_n \approx \sum_{k=1}^{K} \sum_{l=0}^{L-1} \mathbf{w}_{k,l} h_{k,n-l}. \tag{1.39}$$

This model assumes that all frames of a given patch are weighted by the same activation coefficient: $\mathbf{w}_{k,0}$ is weighted by h_{kn} in time frame n, $\mathbf{w}_{k,1}$ by the same h_{kn} in time frame $n+1$, $\mathbf{w}_{k,2}$ by the same h_{kn} in time frame $n+2$, and so on. The full spectrogram $\mathbf{V}$ is therefore approximated as a weighted sum of the patches $\mathbf{W}_k$.

The set of patches $\mathbf{W}_k$ can be partitioned into source-specific dictionaries of patches, which can be learned using NMF, exemplar selection, or exemplar clustering similarly to above [36, 38, 39]. The patch length L is typically on the order of 100 to 300 ms. Fig. 1.4 illustrates a subset of exemplars learned on speech.

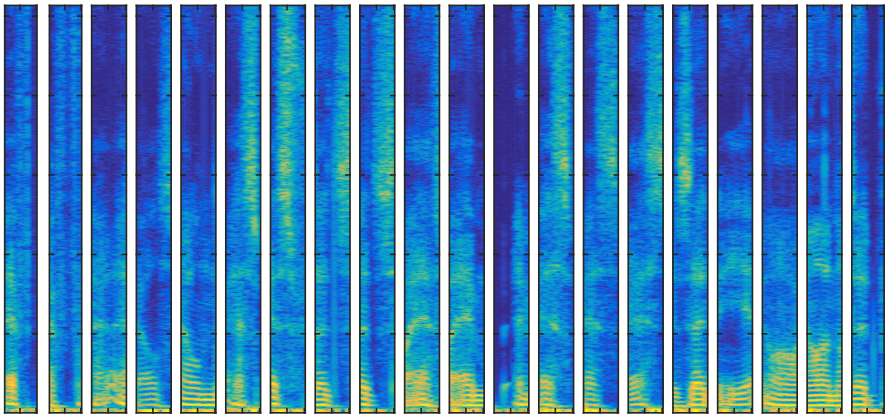

Fig. 1.4 Example convolutive NMF dictionary elements ($\mathbf{W}_k$) learned by random selection of 200 ms exemplars over 500 utterances from a given speaker. Notice how each component represents the spectrogram of a speech phoneme in context.

1.3.1.3 Factoring fine structure and envelope

While supervised NMF makes it possible to account for the characteristics of real audio sources, it is rather constrained and may lead to poor separation when the training and test data exhibit some mismatches. This led to the idea of fixing the source characteristics which remain valid in any circumstances and estimating the other characteristics from the signal to be separated.

is a first step in this direction. The underlying idea is to decompose each dictionary element $\mathbf{w}_k$ in (1.37) as the sum of narrowband spectral patterns $\mathbf{b}_{km}$ weighted by spectral envelope coefficients e_{km}:

$$\mathbf{w}_k = \sum_{m=1}^{M_k} \mathbf{b}_{km} e_{km}. \tag{1.40}$$

The narrowband patterns $\mathbf{b}_{km}$ represent the fine structure of the spectrum and they can be fixed as either smooth or harmonic spectra. In the former case, the patterns can be fixed as smooth narrowband spectra in order to represent a transient or noisy signal with a locally smooth spectrum. In the latter case, each dictionary index k is associated with a given pitch (fundamental frequency) and the corresponding patterns involve a few successive harmonic partials (i.e., spectral peaks at integer multiples of the given fundamental frequency). This model illustrated in Fig. 1.5 is suitable for voiced speech sounds (e.g., vowels) and pitched musical sounds (e.g., violin). The spectral envelope coefficients e_{km} are not fixed, but estimated from the signal to be separated. In other words, this model does not constrain the dictionary elements to match perfectly the training data, but only to follow a certain fine structure.

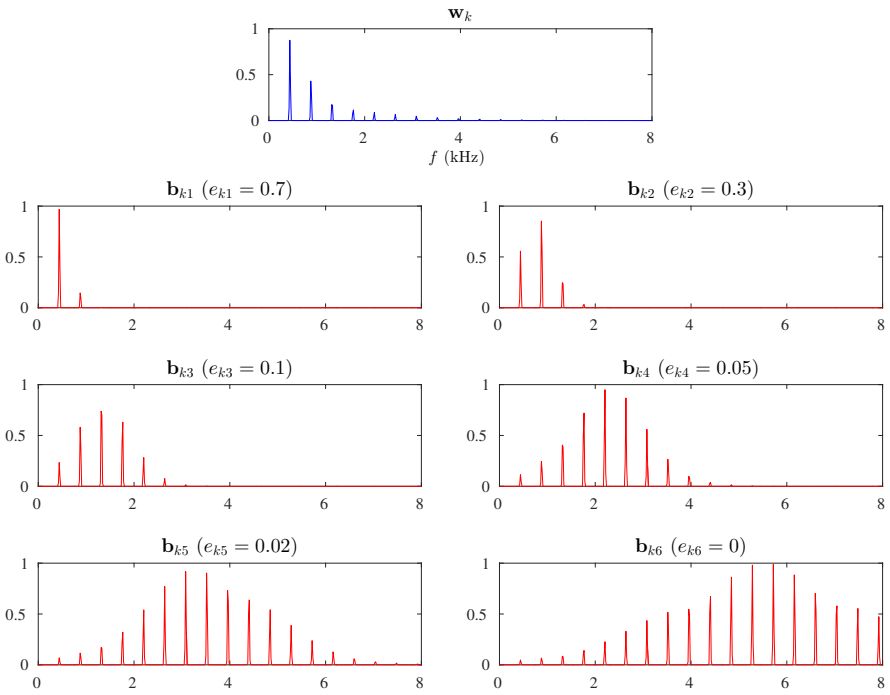

Fig. 1.5 Example narrowband harmonic patterns $\mathbf{b}_{km}$ and resulting dictionary element $\mathbf{w}_k$.

An alternative approach is to factor each dictionary element $\mathbf{w}_k$ into the product of an excitation spectrum and a filter [42]. This so-called *excitation-filter* model adheres with the production phenomena of speech and most musical instruments, where an excitation signal is filtered by the vocal tract or the body of the instrument. The latest evolution in this direction is the multilevel NMF framework of [43], embodied in the Flexible Audio Source Separation Toolbox (FASST)[1]. This framework represents the observed spectrogram as the product of up to eight matrices, which represent the fine structure or the envelope of the excitation or the filter on the time axis or the frequency axis. It makes it possible to incorporate specific knowledge or constraints in a flexible way and it was shown to outperform conventional NMF in [43].

These extensions of NMF are sometimes grouped under the banner of *nonnegative tensor factorisation* (NTF), a generalisation of NMF to multi-dimensional arrays [44]. Due to the linearity of the models, the NTF parameters can be estimated using multiplicative updates similar to the ones for NMF.

[1] `http://bass-db.gforge.inria.fr/fasst/`

1.3.2

1.3.2.1 Sparsity

The original NMF model and the above extensions are well suited for the separation of music sources, which typically involve several overlapping notes. Speech sources, however, consist of a single phoneme at a time. NMF can yield disappointing results on mixtures of speech because it can confuse overlapping phonemes from different speakers vs the same speaker. The latter phenomenon cannot occur due to the physical constraints of speech production, but it is possible according to the model. In order to improve the modelling of speech sources, sparsity constraints must be set on the activation matrix $\mathbf{H}$ [45].

Sparsity signifies that most activation coefficients are very small, and only a small proportion is large. Therefore, it enforces the fact that a single dictionary element predominates in each time frame, and the other dictionary elements are little activated. Sparsity constraints are typically implemented by adding a penalty function to the NMF cost in Section 1.2.1. The ideal penalty function would be the l_0 norm $\|\mathbf{H}\|_0$, that is the number of nonzero entries in $\mathbf{H}$. This norm leads to a combinatorial optimisation problem, though, that is difficult to solve. In practice, the l_1 norm $\|\mathbf{H}\|_1 = \sum_{k=1}^{K}\sum_{n=1}^{N} h_{kn}$ is generally used instead:

$$\arg\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{W},\mathbf{H}) + \mu\|\mathbf{H}\|_1 \tag{1.41}$$

where $\mu > 0$ is a tradeoff parameter.

The penalised cost (1.41) can be minimised w.r.t. $\mathbf{H}$ by adding the constant μ to the denominator of the original multiplicative update [15]:

$$\mathbf{H} \leftarrow \mathbf{H} \circ \frac{\mathbf{W}^T((\mathbf{W}\mathbf{H})^{\circ[\beta-2]} \circ \mathbf{V})}{\mathbf{W}^T(\mathbf{W}\mathbf{H})^{\circ[\beta-1]} + \mu}. \tag{1.42}$$

The greater μ, the sparser the solution. Regarding the dictionary $\mathbf{W}$, the classical update in Section 1.2.3 cannot be used anymore since $\mathbf{W}$ must be normalised in some way in order to avoid scaling indeterminacy, e.g., by assuming each $\mathbf{w}_k$ has a unit l_2 norm $\|\mathbf{w}_k\|_2 = 1$. Rescaling $\mathbf{W}$ a posteriori changes the value of the penalised cost, so that the $\mathbf{W}$ resulting from the classical multiplicative update is not optimal anymore. A multiplicative update accounting for this l_2 norm constraint was proposed in [46, 47]. Alternative sparsity promoting penalties were explored in [48, 49].

1.3.2.2 Group sparsity

Group sparsity is an extension of the concept of sparsity, which enforces simultaneous activation of several dictionary elements. It has been used for two purposes: to automatically group the dictionary elements corresponding to a given phoneme, note or source, in the case when each phoneme, note or source is represented by

multiple dictionary elements [50], and to automatically find which sources are active among a pre-specified set of speakers or musical instruments, when the number of sources and the identity of the active sources are unknown [51].

In the latter case, the full dictionary $\mathbf{W}$ can be partitioned into several source-specific dictionaries $\mathbf{W}_j$ as in Section 1.3.1.1. Group sparsity means that, if source j is inactive, all entries of the corresponding activation matrix $\mathbf{H}_j$ must be estimated as 0. This behaviour can be enforced by using the mixed $l_{1,2}$ norm as a penalty term:

$$\arg\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{W},\mathbf{H}) + \mu \sum_{j=1}^{J} \|\mathbf{H}_j\|_2 \tag{1.43}$$

where the l_2 norm is defined by $\|\mathbf{H}_j\|_2 = (\sum_{k=1}^{K}\sum_{n=1}^{N} h_{jkn}^2)^{1/2}$ and $\mu > 0$ is a tradeoff parameter. Many variants of this penalty can be designed to favour specific activation patterns. For instance, the penalty $\sum_{j=1}^{J}\sum_{n=1}^{N} \|\mathbf{h}_{jn}\|_2$ favours sparsity both over the sources and over time, but all the dictionary elements corresponding to a given source can be activated at a given time. Alternative group sparsity promoting penalties were explored, for instance in [50].

1.3.2.3 Temporal dynamics

Another family of NMF models aim to model the dynamics of the activation coefficients over time. The simplest such models account for the temporal smoothness (a.k.a. continuity) of the activation coefficients by constraining the value of h_{kn} given $h_{k,n-1}$ using a suitable penalty function. In [45], the following penalised cost function was proposed:

$$\arg\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{W},\mathbf{H}) + \sum_{k=1}^{K} \mu_k \sum_{n=2}^{N} (h_{kn} - h_{k,n-1})^2. \tag{1.44}$$

Assuming that μ_k is constant, this penalised cost can be minimised w.r.t. $\mathbf{H}$ by the following multiplicative update inspired from [45]:

$$\mathbf{H} \leftarrow \mathbf{H} \circ \frac{\mathbf{W}^T((\mathbf{W}\mathbf{H})^{\circ[\beta-2]} \circ \mathbf{V}) + 2\mathbf{M} \circ (\overrightarrow{\mathbf{H}} + \overleftarrow{\mathbf{H}})}{\mathbf{W}^T(\mathbf{W}\mathbf{H})^{\circ[\beta-1]} + 2\mathbf{M} \circ (\mathbf{H} + \overleftrightarrow{\mathbf{H}})} \tag{1.45}$$

where

$$\mathbf{M} = \begin{pmatrix} \mu_1 & \cdots & \mu_1 \\ \mu_2 & \cdots & \mu_2 \\ \vdots & \ddots & \vdots \\ \mu_K & \cdots & \mu_K \end{pmatrix} \tag{1.46}$$

$$\overrightarrow{\mathbf{H}} = \begin{pmatrix} 0 & h_{11} & h_{12} & \cdots & h_{1,N-1} \\ 0 & h_{21} & h_{22} & \cdots & h_{2,N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & h_{K1} & h_{K2} & \cdots & h_{K,N-1} \end{pmatrix} \tag{1.47}$$

$$\overleftarrow{\mathbf{H}} = \begin{pmatrix} h_{12} & h_{13} & \cdots & h_{1N} & 0 \\ h_{22} & h_{23} & \cdots & h_{2N} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{K2} & h_{K3} & \cdots & h_{KN} & 0 \end{pmatrix} \tag{1.48}$$

$$\overleftrightarrow{\mathbf{H}} = \begin{pmatrix} 0 & h_{12} & \cdots & h_{1,N-1} & 0 \\ 0 & h_{22} & \cdots & h_{2,N-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & h_{K2} & \cdots & h_{K,N-1} & 0 \end{pmatrix}. \tag{1.49}$$

The impact of μ_k on the resulting activation coefficients is illustrated in Fig. 1.6. The greater μ_k, the smoother the coefficients. Regarding the dictionary $\mathbf{W}$, once again, a normalisation constraint is required which results in a modified update compared to the one in Section 1.2.3. Alternative probabilistically motivated smoothness penalties were proposed in [52, 53].

Building upon this idea, nonnegative continuous-state [54] and discrete-state [34, 55] dynamical models have also been investigated. The latter often limit the number of active dictionary elements at a time and they can be seen as imposing a form of group sparsity. These models account not only for the continuity of the activations, if relevant, but also for typical activation patterns over time due to, e.g., the attack-sustain-decay structure of musical notes or the sequences of phonemes composing common words. For a survey of dynamical NMF models, see [30].

1.3.3

While the above methods incorporate general knowledge about speech and music sources, a number of authors have investigated user-guided NMF methods that incorporate specific information about the sources in a given mixture signal. Existing methods can be broadly categorised according to the nature of this information.

A first category of methods exploit information about the activation patterns of the sources. This information is provided by the user based on listening to the original signal or the separated signals and visualising the waveform or the spectrogram. Given the time intervals when each source is inactive, the corresponding activation

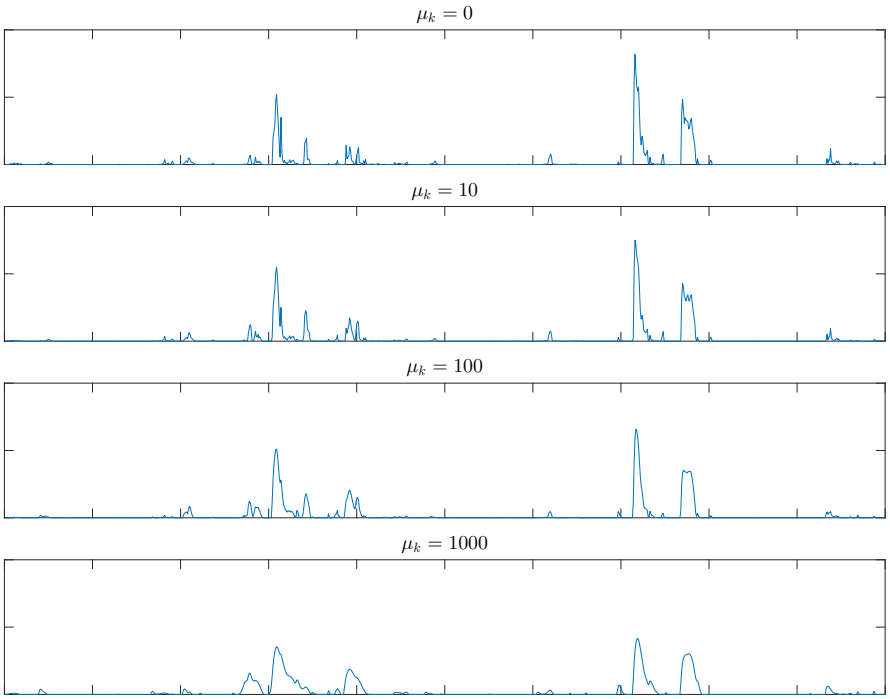

Fig. 1.6 Activation coefficients h_{kn} estimated for one dictionary element k in a music signal for $\beta = 0$ and different values of the smoothness tradeoff parameter μ_k in (1.45).

coefficients can be fixed to 0, which improves the estimation of the dictionary and the activation coefficients in the other time intervals [56]. In [57], a more advanced method is proposed by which the user can tag a given time-frequency region as active, inactive, or well-separated. The graphical user interface is shown in Fig. 1.7. This information is then iteratively exploited in order to refine the source estimates at each iteration. This method was shown to be effective even without using any isolated training data.

A second category of user-guided methods rely on a (partial) transcription of the signal, that can take the form of a fundamental frequency curve [58], a musical score [59], or the speech transcription. This information can be used to restrict the set of active atoms at a given time, in a similar way as group sparsity except that the set of active atoms is known in advance.

Finally, a third category of methods rely on a reference signal for some or all of the sources to be separated. The user can generate reference signals signal by humming the melody [60] or uttering the same sentence [61]. Reference signals can also be obtained by looking for additional data, e.g., the soundtrack of the same film in a different language, the multitrack cover version of a song, additional data corre-

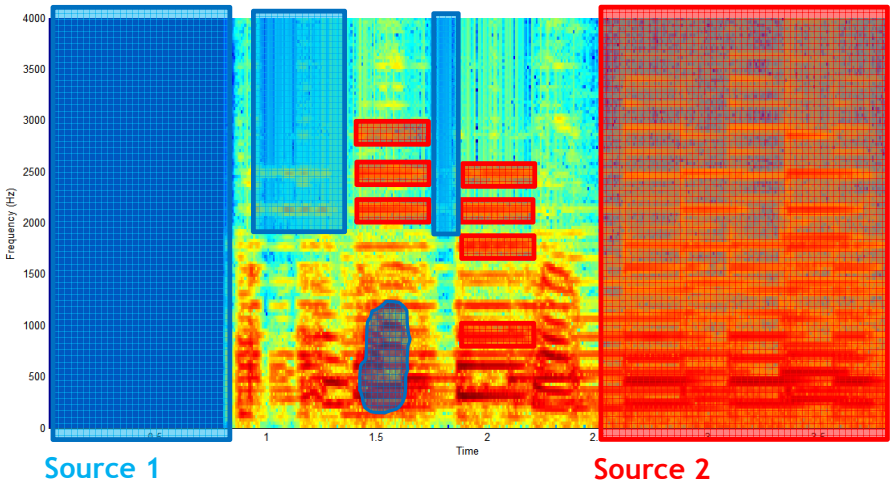

Fig. 1.7 Graphical user interface for user annotation. Piano is labelled as active (resp. inactive) in the red (resp. blue) regions.

sponding to the same speaker or the same musical instrument, or repeated signals (e.g., jingles, background music) in large audio archives [62].

Many user-guided NMF methods can be expressed under the general framework of nonnegative matrix partial co-factorisation (NMPcF), which aims to jointly factor several input matrices into several factor matrices, some of which are shared [63, 64]. For instance, in the case of score-guided or reference-guided separation, the spectrogram to be separated and the score or the reference can be jointly factored using different dictionaries but the same activation matrix.

1.4 Conclusions

In this chapter, we have shown that NMF is a powerful approach for audio source separation. Starting from a simple unsupervised formulation, it makes it possible to incorporate additional information about the sources in a principled optimisation framework. In comparison with deep neural network (DNN) based separation, which has recently attracted a lot of interest, NMF-based separation remains competitive in the situations when the amount of data is medium or small, or user guidance is available. These two situations are hardly handled by DNNs today, due to the need for a large amount of training data and the difficulty of retraining or adapting the DNN at test time based on user feedback. It therefore comes as no surprise that NMF is still the subject of much research today. Most of this research concentrates on overcoming the fundamental limitation of NMF, namely the fact that it models

spectro-temporal magnitude or power only, and enabling it to account for phase. For an in-depth discussion of this and other perspectives, see [65].

On a final note, some aspects of NMF for audio signal processing are also covered in other chapters of the present book (Chapters ?, ? , ?) and in Chapters 8, 9 & 16 of [65].

Standard distributions

Poisson

$$Po(x|\lambda) = \exp(-\lambda)\,\frac{\lambda^x}{x!}, \quad x \in \{0,1,\ldots,\infty\} \tag{1.50}$$

Multinomial

$$M(\mathbf{x}|N,\mathbf{p}) = \frac{N!}{x_1!\ldots x_K!} p_1^{x_1}\ldots p_K^{x_K}, \quad x_k \in \{0,\ldots,N\}, \sum_k x_k = N \tag{1.51}$$

Circular complex normal distribution

$$N_c\,(x|\mu,\Sigma) = |\pi\,\Sigma|^{-1}\,\exp -(x-\mu)^H\,\Sigma^{-1}\,(x-\mu), \quad x \in \mathbb{C}^F \tag{1.52}$$

Gamma

$$G(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,x^{\alpha-1}\,\exp(-\beta\,x), \quad x \geq 0 \tag{1.53}$$

References

[1] C. J. C. Burges, "Dimension reduction: A guided tour," *Foundations and Trends in Machine Learning*, vol. 2, no. 4, pp. 275–365, 2009.

[2] P. Comon, "Independent component analysis, a new concept ?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[3] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[4] M. Aharon, M. Elad, and A. Bruckstein, "K -SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[5] P. Paatero and U. Tapper, "Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 111–126, 1994.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[7] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

[8] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computers*, vol. 42, no. 8, pp. 30–37, 2009.

[9] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero, "Nonlinear unmixing of hyperspectral images: Models and algorithms," *IEEE Signal Proccessing Magazine*, vol. 31, no. 1, pp. 89–94, Jan. 2014.

[10] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2003.

[11] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[12] A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi, "Non-negative matrix factorization with α-divergence," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1433–1440, July 2008.

[13] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, Charleston SC, USA, Mar. 2006, pp. 32–39.

[14] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, 2007.

[15] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011.

[16] I. S. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," in *Advances in Neural Information Processing Systems (NIPS)*, 2005.

[17] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, Sep. 1998.

[18] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," Institute of Statistical Mathematics, Tech. Rep., June 2001, research Memo. 802.

[19] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Proc. Irish Signals and Systems Conference*, 2009.

[20] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model non stationary audio events," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 445–448.

[21] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, pp. 528 – 537, 2010.

[22] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the beta-divergence," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592 – 1605, July 2013.

[23] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *Proc. 17th European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, Aug. 2009, pp. 1913–1917.

[24] J. F. Canny, "GaP: A factor model for discrete data," in *Proc. ACM International Conference on Research and Development of Information Retrieval (SIGIR)*, 2004, pp. 122–129.

[25] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, no. Article ID 785152, p. 17 pages, 2009, doi:10.1155/2009/785152.

[26] P. Smaragdis, B. Raj, and M. V. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *NIPS workshop on Advances in models for acoustic processing*, 2006.

[27] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[28] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sep. 2012.

[29] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, pp. 30 – 37, 2004.

[30] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, May 2014.

[31] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proc. International Computer Music Conference (ICMC)*, 2003, pp. 231–234.

[32] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 337–344.

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

[34] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," in *Proc. International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, 2004, pp. 327–334.

[35] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi- supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 17–20.

[36] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[37] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Proc. Neural Information Processing Systems (NIPS)*, 2009, pp. 1705–1713.

[38] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[39] T. Virtanen, J. Gemmeke, and B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2277–2289, Nov. 2013.

[40] P. D. O'Grady and B. A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, vol. 72, no. 1-3, pp. 88 – 101, 2008.

[41] W. Wang, A. Cichocki, and J. A. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared euclidean distance," *IEEE Trans. Signal Processing*, vol. 57, no. 7, pp. 2858–2864, July 2009.

[42] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[43] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[44] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, vol. 2008, 2008, article ID 872425.

[45] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[46] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proc. IEEE International Joint Conference on Neural Networks*, 2004, pp. 2529–2533.

[47] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse NMF – half-baked or well done?" Mitsubishi Electric Research Laboratories (MERL), Tech. Rep. TR2015-023, Mar. 2015.

[48] C. Joder, F. Weninger, D. Virette, and B. Schuller, "A comparative study on sparsity penalties for NMF-based speech separation: Beyond Lp-norms," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 858–862.

[49] Y. Mitsui, D. Kitamura, S. Takamichi, N. Ono, and H. Saruwatari, "Blind source separation based on independent low-rank matrix analysis with sparse regularization for time-series activity," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[50] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[51] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[52] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[53] O. Dikmen and A. T. Cemgil, "Gamma Markov random fields for audio source modeling," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 589–601, 2010.

[54] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 3158–3162.

[55] G. Mysore and M. Sahani, "Variational inference in non-negative factorial hidden Markov models for efficient audio source separation," in *Proc. International Conference on Machine Learning (ICML)*, 2012, pp. 1887–1894.

[56] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, May 2011, pp. 257–260.

[57] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization,"

in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 1567–1571.

[58] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected F0 track," in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012, pp. 438–445.

[59] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, May 2014.

[60] P. Smaragdis and G. J. Mysore, "Separation by humming : User-guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 69–72.

[61] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[62] N. Souviraà-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Multi-channel audio source separation using multiple deformed references," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1775–1787, 2015.

[63] Y. K. Yılmaz, A. T. Cemgil, and U. Şimşekli, "Generalized coupled tensor factorization," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[64] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorization," in *IEEE Trans. Signal Processing*, vol. 62, no. 22, pp. 5940–5949, Nov. 2014.

[65] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*, Wiley, 2017.