An introduction to multichannel NMF for audio source separation

Alexey Ozerov, Cédric Févotte, Emmanuel Vincent

Chapter 1
An introduction to multichannel NMF for audio source separation

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

Abstract

This chapter introduces multichannel nonnegative matrix factorization (NMF) methods for audio source separation. All the methods and some of their extensions are introduced within a more general local Gaussian modeling (LGM) framework. These methods are very attractive since allow combining spatial and spectral cues in a joint and principal way, but also are natural extensions and generalizations of many single-channel NMF-based methods to the multichannel case. The chapter introduces the spectral (NMF-based) and spatial models, as well as the way to combine them within the LGM framework. Model estimation criteria and algorithms are described as well, while going deeper into details of some of them.

1.1 Introduction

Nonnegative matrix factorisation (NMF) [1] is a dimensionality reduction technique that consists in approximating a nonnegative data matrix (a matrix with nonnegative entries) as a product of two nonnegative matrices of lower rank than the initial data matrix. This also can be viewed as an approximation of data matrix as a sum of few rank-1 nonnegative matrices. It was first successfully applied for single-channel source separation [2], where the nonnegative matrix of magnitude or power spectrogram is decomposed, and became a state of the art reference. The success of this method is mainly due to universality of this quite simple modeling (it is applicable to various types of audio sources including speech [3, 4], music [2, 5], environmental sounds [6], etc.) and due to the flexibility of this modeling allowing adding various constraints to it, such as for example harmonicity of spectral patterns [7],

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

smoothness of their activation coefficients [2, 5], pre-trained spectral patterns [8, 9], etc.

Given the success of the NMF for single-channel source separation, there were several attempts to extend it to the case of multichannel source separation. Earlier ideas were relying on stacking magnitude or power spectrograms of all channels into a 3-valence nonnegative tensor and decomposing it with nonnegative tensor factorisation (NTF) methods [10] or other NTF-like nonnegative structured approximations [11, 12]. This gave some interesting results. However, since only nonnegative power spectrograms are involved, such approaches rely only on the amplitude information, while completely discarding the phases of the short time Fourier transforms (STFTs). In other words, these approaches do not allow exploiting the interchannel phase differences (IPDs), but only the interchannel level differences (ILDs). However, the IPDs may be very important for multichannel source separation, and they are indeed exploited by several clustering-based methods [13, 14]. Using IPDs becomes even more critical for the far-field case (i.e., when the distances between the microphones are much smaller than the distances between the sources and microphones), where the information carried by the ILDs becomes almost non-discriminating.

It is clear that a fully nonnegative (e.g., NTF-like) modeling is unable to model jointly source power spectrograms, ILDs and IPDs, since the phase information is discarded in the nonnegative tensor of multichannel mixture power spectrograms. As such, it was proposed to resort to a semi-nonnegative modeling [12, 15, 8, 16, 17], where the latent source power spectrograms are modeled with NMF [12, 8] or NTF [15, 16, 17], while the mixing system is modeled differently, not with a nonnegative model. This modeling, often referred to as *multichannel NMF* [12] or *multichannel NTF* [15] [1] depending on the model of the source power spectrograms, is usually achieved via a Gaussian probabilistic modeling applied directly to the complex-valued STFTs of all channels.

The multichannel NMF modeling treats the complex-valued STFT coefficients as realizations of zero-mean circular complex-valued Gaussian random variables with structured variances (via NMF) and covariances. This leads to the fact that this modeling reduces to Itakura Saito (IS) NMF in the single channel case (see Chapter *Févotte et al*), thus being its natural extension to the multichannel case. Moreover, it allows integrating many other NMF-like models (see Chapter *Févotte et al* and [8]) in an easy and flexible manner. Finally, it combines both spectral and spatial (including ILDs and IPDs) cues within a unified framework. When one of these two cues does not allow separating the sources efficiently, the algorithm relies on the other cue, and vice versa. In our opinion the multichannel NMF is one of the first attempts of combining these two cues in a systematic and principal way.

[1] Throughout the chapter we will generally refer to all these methods as multichannel NMF, while precising when we are speaking about multichannel NTF.

1.2 Local Gaussian model

Multichannel NMF can be formulated as based on a so-called *local Gaussian model (LGM)* that is more general itself (than the multichannel NMF) and allows modeling and combining spatial and spectral cues in a systematic way. In a most general manner the LGM may be formulated as follows. Let us first assume that we deal with a multichannel (I-channel) mixture of J sources to be separated. Assuming all the signals are converted into the STFT domain, this can be written as

$$\mathbf{x}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{jfn}, \tag{1.1}$$

where $\mathbf{x}_{fn} = \left[x_{1,fn}, \ldots, x_{I,fn}\right]^T \in \mathbb{C}^I$ and $\mathbf{y}_{jfn} = \left[y_{1,jfn}, \ldots, x_{I,jfn}\right]^T \in \mathbb{C}^I$ ($j = 1, \ldots, J$) are the channel-wise vectors of STFT coefficients of the mixture and of the j-th source *spatial image* [2], respectively; and $f = 1, \ldots, F$ and $n = 1, \ldots, N$ are the frequency and time indices, respectively. Given the above-introduced notations, the LGM modeling [18] assumes that each source image (I-length complex-valued vector $\mathbf{y}_{jfn}$) is modeled as a zero-mean circular complex Gaussian random vector as follows

$$\mathbf{y}_{jfn} \sim \mathcal{N}_c\left(0, \mathbf{R}_{jfn} v_{jfn}\right), \tag{1.2}$$

where the complex-valued covariance matrix is positive definite Hermitian, and it is composed of two factors:

- a *spatial covariance* $\mathbf{R}_{jfn} \in \mathbb{C}^{I \times I}$ representing the spatial characteristics of the j-th source image at the time-frequency (TF) point (f, n), and
- a *spectral variance* $\mathbf{v}_{jfn} \in \mathbb{R}$ representing the spectral characteristics of the j-th source image at the TF point (f, n).

Given the model parameters, i.e., the spatial covariances $\mathbf{R}_{jfn}$ and the spectral variances $\mathbf{v}_{jfn}$, the random vectors $\mathbf{y}_{jfn}$ in (1.2) are also assumed mutually independent in time, frequency and between sources. Note that the LGM modeling was not proposed in [18] for the first time, indeed, its variants were already considered in [19, 20]. However, the formulation from [18] is quite general to cover all the cases, that is why we have chosen here this formulation.

Given the multichannel mixing equation and the above independence assumptions, the mixture STFT coefficients may be shown distributed as

$$\mathbf{x}_{fn} \sim \mathcal{N}_c\left(0, \sum_{j=1}^{J} \mathbf{R}_{jfn} v_{jfn}\right). \tag{1.3}$$

[2] The spatial image of a source means not the source signal itself, but its contribution into the I-channel mixture.

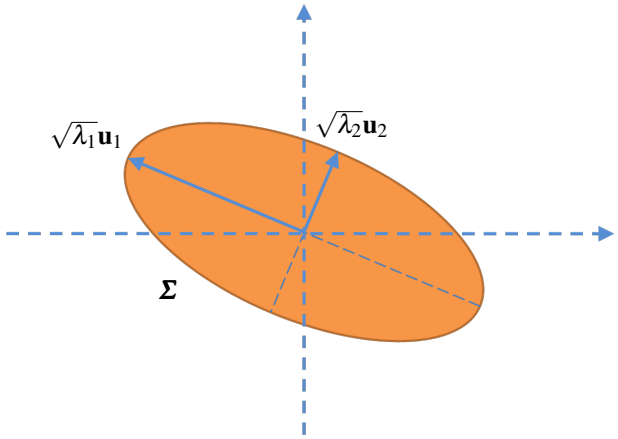

Fig. 1.1 An illustration of a spatial covariance matrix $\mathbf{R}_{jfn}$ in the 2-channel case ($I = 2$). While dropping the indices j, f and n, the covariance matrix eigendecomposition may be written as $\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H$, with $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2]$, $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{C}^2$ being the eigenvectors and $\boldsymbol{\Lambda} = \mathrm{diag}([\lambda_1, \lambda_2])$, $\lambda_1, \lambda_2 \in \mathbb{R}_+$ being the eigenvalues. This illustration is not fully complete, since a 2D complex-valued covariance matrix is represented on a 2D real plane.

The model parameters are usually estimated in the maximum likelihood (ML) sense from the observed mixture $\mathbf{X} = \{x_{ifn}\}_{i,f,n}$. However, a direct ML estimation of parameters under the modeling (1.3) would lead to the data overfitting, since the number of scalar parameters exceeds the number of the mixture STFT coeffitients. As such, various constraints are applied to both spectral variances and spatial covariances, as it is presented in detail in Sections 1.3 and 1.4 respectively. In the case of multichannel NMF we address in this chapter, the spectral variances are usually represented by low-rank nonnegative matrices or tensors. However, other approaches consider different models (e.g., such as composite autoregressive models [21], source-excitation models [8] or hidden Markov models [22]) to structure the spectral variances, that is why the LGM modeling is more general than the multichannel NMF. As it is discussed in Section 1.4 below, spectral covariances are usually not modeled with fully nonnegative structures. This is the reason why we are speaking about semi-nonnegative modeling in the introduction.

For the sake of better understanding, we now give an interpretation to the spatial covariance matrix $\mathbf{R}_{jfn}$, and relate it to the methods used for multichannel audio compression. For the sake of simplicity and also since most of audio recording are stereo (i.e., two channel mixtures), we consider the case of $I = 2$. The spatial covariance matrix $\mathbf{R}_{jfn}$ is in general a full-rank positive definite Hermitian complex-valued matrix. An example of a spatial covariance matrix is represented on Figure 1.1. Note that this is a rather "fake" (or incomplete) representation, since it is difficult to represent a 2-dimensional complex-valued covariance matrix on a 2-dimentional real plane.

Since the spatial covariance matrix $\mathbf{R}_{jfn}$ is complex-valued Hermitian, it can be easily shown that in the 2-dimensional case we consider here it is uniquely encoded by only four real scalars. Indeed, its 2 diagonal entries are real and the 2 complex-valued off-diagonal entries are conjugate. These four real-valued parameters may be uniquely converted into the following, in a sense more meaningful, real-valued parameters:

- Loudness [3],
- ILD,
- IPD,
- Diffuseness that can be also replaced by interchannel coherence (IC) [23].

It is worth to note that the last three spatial parameters (ILD, IPD and IC) are also used for parametric coding of stereo audio [23]. This is somehow expected, indeed, the models that are suitable for compression should be also suitable for sources separation, since in both cases the models tend to reduce the redundancy in the signal.

Finally, let us also stress that the LGM modeling seems more general (and thanks to Gaussian formulation more principal) than blind source separation (BSS) approaches based on ILD/IPD clustering [13, 24]. Indeed, the diffuseness or IC is not taken at all into account within the latter approaches.

1.3 Spectral models

In this section we present and discuss spectral models used within various multichannel NMF approaches. These models include NMF models, NTF models and their extensions.

1.3.1 NMF modeling of each source

NMF modeling of each source, which is usually referred to as mutichannel NMF, consists in structuring the source variances v_{jfn} in (1.2) with NMF structure as in the single-channel NMF case (see Chapter *Févotte et al.*):

$$v_{jfn} = \sum_{k=1}^{K_j} w_{jfk} h_{jkn}, \tag{1.4}$$

where the source-dependent K_j is usually smaller than both F and N, and w_{jfk} and h_{jkn} are all nonnegative. By introducing nonnegative matrices (i.e., matrices with

[3] Due to the scale ambiguity between $\mathbf{R}_{jfn}$ and v_{jfn} in (1.2), the loudness can be fully attributed to v_{jfn}.

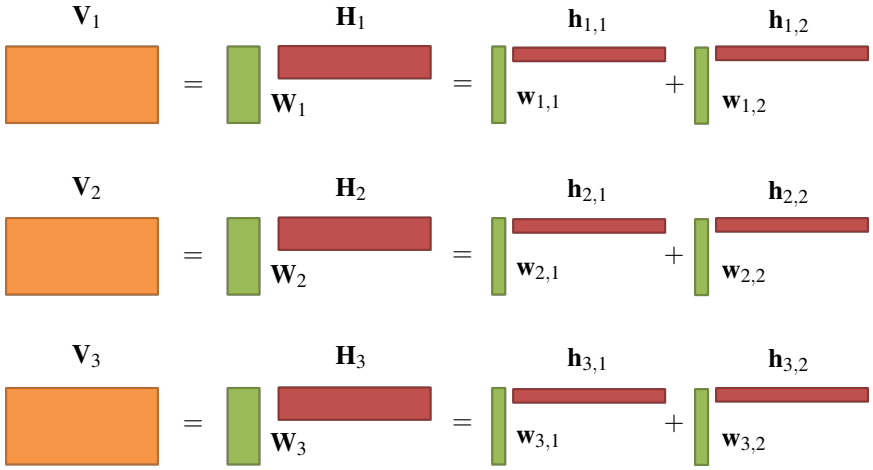

Fig. 1.2 A visualization of spectral models of multichannel NMF. Source variances $\mathbf{V}_j$ of each of J (here $J = 3$) sources are modeled with NMF with K_j (here $K_j = 2$) components, which can be decomposed as a sum of K_j rank-1 matrices ($\mathbf{w}_{j,k}$ and $\mathbf{h}_{j,k}$ are the columns and the lines of matrices $\mathbf{W}$ and $\mathbf{H}$, respectively).

nonnegative entries) $\mathbf{V}_j = [v_{jfn}]_{f,n} \in \mathbb{R}_+^{F\times N}$, $\mathbf{W}_j = [w_{jfk}]_{f,k} \in \mathbb{R}_+^{F\times K_j}$, and $\mathbf{H}_j = [h_{jkn}]_{k,n} \in \mathbb{R}_+^{K_j\times N}$, equation (1.4) may be rewritten in a matrix form as:

$$\mathbf{V}_j = \mathbf{W}_j\mathbf{H}_j. \tag{1.5}$$

A visualization of these NMF spectral models is shown on Figure 1.2.

This kind of spectral models in the case of mutichannel source separation were first introduced in [25, 26], though with more sophisticated NMF-like structures suitable for harmonic music instruments and with different optimization criteria than those we discuss in this chapter. Spectral models based on usual NMF, exactly as in (1.5), were proposed in [12], and then extended/re-considered in many other works [27, 15, 8, 16, 17].

A very attractive property of this modeling is that any NMF or NMF-like structure based on the IS divergence, such as for example harmonic NMF [7], smooth NMF [2, 5] or excitation-filter NMF [28] (see also Chapter *Févotte et al.*) may be incorporated easily and in a systematic manner within the framework. This was remarked and addressed in [8], where a general source separation framework allowing specifying various spectral and spatial models for each individual source is proposed. The latter research work is supplied with a software called Flexible Audio Source Separation Toolbox (FASST) that implements all these possible model variants in a flexible way. Finally, let us note that many informed or user-assisted/guided audio source separation approaches were extended to the multichannel case within the same paradigm [15, 29].

1.3.2 Joint NTF modeling of all sources

One of the shortcomings of the multichannel NMF modeling presented in Section 1.3.1 is the following. While for single-channel NMF one needs fixing an appropriate number of components K or determining this number automatically, which is not always easy (see, e.g., [30]), in the multichannel NMF, as presented in Sec. 1.3.1, one needs determining not only the total number of components $K = \sum_{j=1}^{J} K_j$, but also the number of components K_j for each source, which may vary from one source to another. To overcome this problem the following idea was introduced in [15], and then extended in other works [16, 17]. It is now assumed that instead of representing each source with an individual NMF $\{\mathbf{W}_j, \mathbf{H}_j\}$ all the sources share the components of the same NMF $\{\mathbf{W}, \mathbf{H}\}$, where $\mathbf{W} = [w_{fk}]_{f,k} \in \mathbb{R}_+^{F\times K}$, and $\mathbf{H} = [h_{kn}]_{k,n} \in \mathbb{R}_+^{K\times N}$. Moreover, in order to specify associations between K NMF components and J sources, a new $(J \times K)$ nonnegative matrix $\mathbf{Q} = [q_{jk}]_{j,k} \in \mathbb{R}_+^{J\times K}$ is introduced, and the source variances v_{jfn} are now structured as:

$$v_{jfn} = \sum_{k=1}^{K} w_{fk} h_{kn} q_{jk}. \tag{1.6}$$

Assuming the columns of $\mathbf{Q}$ are normalized to sum to one (i.e., $\sum_{j=1}^{J} q_{jk} = 1$), which is always possible to achieve thanks to scale ambiguity between the columns of $\mathbf{Q}$ and that of say $\mathbf{W}$ in (1.6), each q_{jk} represents the proportion of association of the component k to the source j.

By denoting with $\mathbf{V} = \{v_{jfn}\}_{j,f,n}$ a 3-valence tensor of source variances, equation (1.6) may be also rewritten in a tensor/vector form as a sum of K rank-1 tensors:

$$\mathbf{V} = \sum_{k=1}^{K} \mathbf{w}_k \circ \mathbf{h}_k^T \circ \mathbf{q}_k, \tag{1.7}$$

where "$\circ$" denotes the tensor outer product, $\mathbf{w}_k$ and $\mathbf{q}_k$ are the k-th columns of matrices $\mathbf{W}$ and $\mathbf{Q}$ respectively, and $\mathbf{h}_k$ is the k-th line of matrix $\mathbf{H}$. The tensor decomposition as in (1.6) and (1.7) is called parallel factor (PARAFAC) or canonical decomposition (CANDECOMP) [31]. A visualization of these NTF spectral models is shown on Figure 1.3.

We here call this model multichannel NTF, as introduced in [15], though some authors [16, 17] continue calling it multichannel NMF. Note also that a fully nonnegative NTF modeling [10, 11, 12] was applied for multichannel audio source separation as well. Those approaches apply an NTF decomposition directly to the nonnegative tensor of power spectrograms of the multichannel mixture, while here it is applied to the latent nonnegative tensor of power spectrograms of the sources, and the overall modeling is not fully nonnegative, as mentioned in the introduction.

One can easily note that the NTF decomposition (1.6) generalizes that of (1.4). Indeed, (1.6) can be reduced to (1.4) by setting for each column of $\mathbf{Q}$ all the values to 0 except one that is set to 1, and by fixing the values of $\mathbf{Q}$. Finally, the multichannel

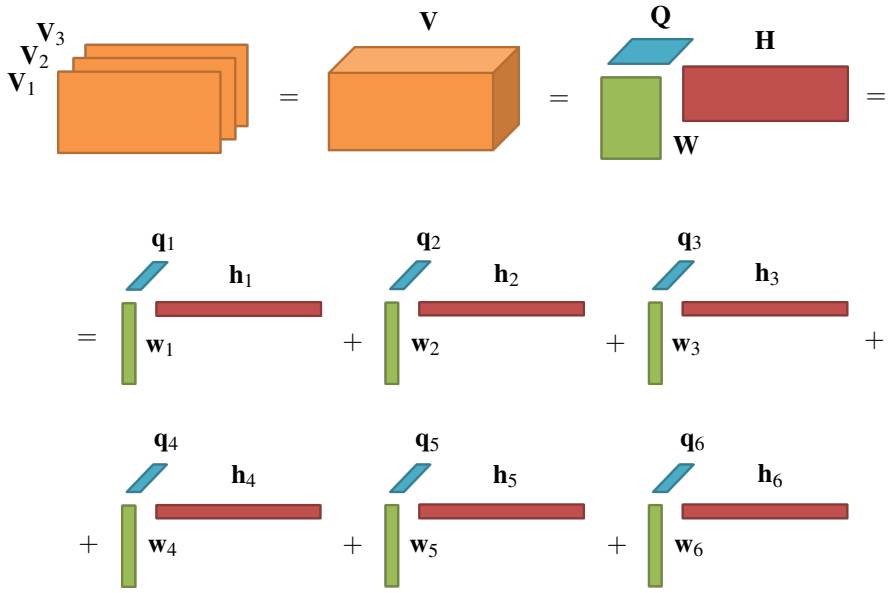

Fig. 1.3 A visualization of spectral models of multichannel NTF. Source variances $\mathbf{V}_j$ are stuck in a common 3-valence tensor $\mathbf{V}$ modeled with PARAFAC model [31] with K (here $K = 6$) components, which can be decomposed as a sum of K rank-1 3-valence tensors.

NTF modeling has the following potential advantages over the multichannel NMF modeling:

- One does not need specifying in advance the number of components K_j for each source, but only the total number of components K. The components are then allocated automatically via the matrix $\mathbf{Q}$, which may be also more optimal than a manual user-specified allocation.
- Some components may be shared between different sources, which means that the modeling is more compact. This happens when there are more than one non-zero entry in one column of matrix $\mathbf{Q}$.

It should be noted however that it is desirable that the matrix $\mathbf{Q}$ is quite sparse, i.e., that there are few components for which there are more than one non-zero entry in the corresponding column of matrix $\mathbf{Q}$. Otherwise, the components are not well allocated between sources, and this may not lead to a good separation result. Thus, it is possibly desirable to add some sparsity-inducing penalty on $\mathbf{Q}$ to the corresponding optimization criterion.

1.4 Spatial models and constraints

Spatial covariance $\mathbf{R}_{jfn}$ might be assumed fully unconstrained, though in that case, as already mentioned in Section 1.2, the parameter estimation would certainly lead to data overfitting, since there are more parameters than observations, i.e., the STFT coefficients in the multichannel mixture. In order to cope with that it is necessary to introduce some constraints on spatial covariances.

First of all, when the sources are static, it is reasonable to assume that the spatial covariances are time-invariant, i.e., $\mathbf{R}_{jfn} = \mathbf{R}_{jf}$ are independent of n. This assumption is made in many approaches [18, 12, 8, 16, 17] and it allows highly reducing the number of free parameters to be estimated. We assume the time-invariant case within this section and the time-varying case will be briefly discussed at the end.

On top of the time-invariance, additional constraints may be introduced as well, and most often it is achieved either by imposing some particular structure or via probabilistic priors.

The early works [19, 20, 12] constraint the spatial covariance $\mathbf{R}_{jf}$ further and assume that the rank of the matrix is one, which is refereed to as *rank-1 spatial covariance*. This was introduced based on the following reasoning. Let us assume that the mixture (1.2) is a convolutive mixture of J point sources. In that case the spatial images $\mathbf{y}_{jfn}$ in (1.2) may be approximated as [32]

$$\mathbf{y}_{jfn} = \mathbf{a}_{jf} s_{jfn}, \tag{1.8}$$

where $s_{jfn} \in \mathbb{C}$ are the STFT coefficients of the point sources and $\mathbf{a}_{jf} = \left[a_{1jf}, \ldots, a_{Ijf}\right]^T \in \mathbb{C}^I$ are the channel-wise vectors of discrete Fourier transforms (DFTs) of the impulse responses of the convolutive mixing filters. The equality in (1.8) holds indeed only approximately and becomes more and more accurate when the sizes of the mixing filters impulse responses are comparable or smaller than the length of the STFT analysis window [32]. This approximation is referred to as *narrowband approximation*. Assuming now that each source STFT coefficient s_{jfn} follows a zero-mean Gaussian distribution with variance v_{jfn}, one can easily show that source images $\mathbf{y}_{jfn}$ are distributed as in (1.2) with

$$\mathbf{R}_{jf} = \mathbf{a}_{jf} \mathbf{a}_{jf}^H. \tag{1.9}$$

We see that the spatial covariance $\mathbf{R}_{jf}$ in (1.9) is indeed a rank-1 matrix.

It was proposed in [18] not to constraint the spatial covariance $\mathbf{R}_{jf}$ or to parametrize it in a different way (see [18] for details), but in both cases so as the matrix remains full rank. This modeling, refereed to as *full rank spatial covariance*, allows to go beyond the limits of the narrowband approximation (1.8), thus it is more suitable than the rank-1 model in case of long reverberation times. It may be also more suitable in case when the point sources assumption is not fully verified. Indeed, as explained in Section 1.7.2 below, modeling a source image with a full rank model can be recast as a sum of I point sources with different rank-1 spatial covariances and shared spectral variance.

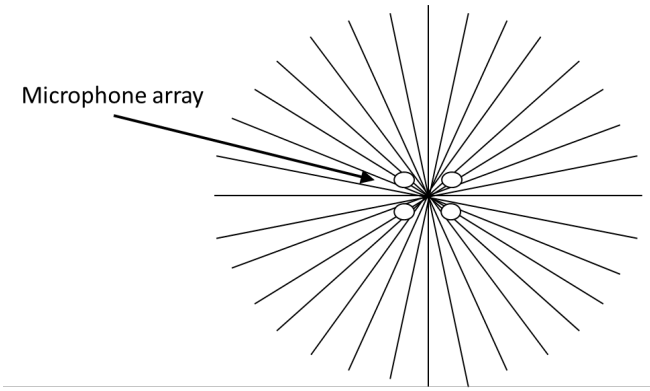

Fig. 1.4 Example of a set of predefined directions in 2D plane for a given microphone array.

Another approach [17] consists in assuming that the spatial covariance is a weighted sum of so-called *direction of arrival (DOA) kernels* that are rank-1 spatial covariances modeling plane waves coming from several predefined directions. These directions may be specified in 2D plane or in 3D space (see Fig. 1.4 for a 2D example). Rank-1 DOA kernels corresponding to these directions θ_l ($l = 1, \ldots, L$) are then defined as

$$\mathbf{K}_{fl} = \mathbf{d}(f, \theta_l)\mathbf{d}(f, \theta_l)^H \tag{1.10}$$

with $\mathbf{d}(f, \theta_l)$ being a *relative steering vector* for the direction θ_l defined as

$$\mathbf{d}(f, \theta_l) = \left[1, e^{-2\pi\tau_{2,1}(\theta_l)\nu_f/c}, \ldots, e^{-2\pi\tau_{I,1}(\theta_l)\nu_f/c}\right]^T, \tag{1.11}$$

where c is the speed of the sound (343 m/s), ν_f is the frequency (in Hz) corresponding to the frequency bin f, and $\tau_{i,i'}(\theta_l)$ is the time difference of arrival (TDOA) (in seconds) between microphones i and i' from the direction θ_l. Note that this relative steering vector is defined without taking into account the ILDs, but only IPDs (see [33] for a definition taking as well into account ILDs). Finally, the spatial covariance is defined as a weighted sum of DOA kernels $\mathbf{K}_{fl}$ from (1.10) as

$$\mathbf{R}_{jf} = \sum_{l=1}^{L} z_{jl}\mathbf{K}_{fl}, \tag{1.12}$$

with z_{jl} being nonnegative weights.

If the DOAs of all or of some sources are known to some extend, it is possible to introduce this information for example via prior distributions on the spatial covariances. In [34] those priors are defined via inverse Wishart distributions as follows

$$p\left(\mathbf{R}_{jf}|\boldsymbol{\Psi}_{jf},m\right)=\frac{|\boldsymbol{\Psi}_{jf}|^{m}|\mathbf{R}_{jf}|^{-(m+I)}e^{-\mathrm{tr}\left[\boldsymbol{\Psi}_{jf}\mathbf{R}_{jf}^{-1}\right]}}{\pi^{I(I-1)/2}\prod_{i=1}^{I}\Gamma(m-i+1)},\tag{1.13}$$

with

$$\boldsymbol{\Psi}_{jf}=(m-I)\left(\mathbf{d}(f,\theta_l)\mathbf{d}(f,\theta_l)^{H}+\sigma_{\mathrm{rev}}^{2}\boldsymbol{\Omega}_{f}\right),\tag{1.14}$$

where $\mathbf{d}(f,\theta_l)$ is a steering vector which may be defined as in (1.11), $\boldsymbol{\Omega}_f=\left[\sin(2\pi\nu_f q_{ii'}/c)/(2\pi\nu_f q_{ii'}/c)\right]_{ii'}$ is a matrix modeling reverberation part (i.e., non-direct part) of the impulse response, and σ_{rev}^2 is a positive constant depending on the amount of reverberation as compared to the direct part of impulse response.

There are also other models that do not fall into the LGM framework as formulated here. These models include for example multichannel high-resolution NMF (HR-NMF) [35] or a method where the source variance prior parametrization is factorized by NMF [36].

Finally, several approaches [37, 38, 39] address time-varying case, where $\mathbf{R}_{jfn}$ is not independent any more on n, though still constrained in different ways.

1.5 Main steps and sources estimation

Let us denote by $\boldsymbol{\theta}=\{\mathbf{R}_{jfn},v_{jfn}\}_{j,f,n}$ the whole set of model parameters, assuming some constraints from those overviewed in Sections 1.3 and 1.4 hold. Given a model $\boldsymbol{\theta}$ specified and an estimation criterion (see Section 1.6 below) chosen, most of LGM-based approaches are based on the following main steps:

1. The STFT $\mathbf{X}$ of the multichannel mixture signal is computed.
2. The model is estimated with an algorithm (see Section 1.7 below) optimizing the chosen criterion.
3. The source images are estimated in the STFT domain via Wiener filtering as:

$$\hat{\mathbf{y}}_{jfn}=\mathbf{R}_{jfn}v_{jfn}\left[\sum_{j=1}^{J}\mathbf{R}_{jfn}v_{jfn}\right]^{-1}\mathbf{x}_{fn},\tag{1.15}$$

   where $\mathbf{R}_{jfn}$ and v_{jfn} are the spatial covariances and spectral variances as specified in (1.2).
4. The source images in time domain are then reconstructed by applying the inverse STFT to $\widehat{\mathbf{Y}}=\{\hat{\mathbf{y}}_{jfn}\}_{j,f,n}$.

In the online approaches [40, 41], where the separation must be performed for every new frame, the same steps are repeated for each frame and the model estimation algorithm is modified so as to update the model parameters in an incremental and causal (i.e., only the passed and current frames are used) manner.

1.6 Model estimation criteria

In order to estimate the model parameters $\boldsymbol{\theta}$ from the observed data, i.e., from the STFT of the multichannel mixture signal $\mathbf{X}$, one needs specifying a model estimation criterion.

1.6.1 Maximum likelihood

One of the most popular choices for model estimation is the maximum likelihood (ML) criterion that writes

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}'} p(\mathbf{X}|\boldsymbol{\theta}'). \tag{1.16}$$

In the case of LGM modeling (1.2) this criterion can be shown [16] equivalent to minimizing the following cost function:

$$C_{\mathrm{IS}}(\boldsymbol{\theta}) = \sum_{f,n=1}^{F,N} \mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\right) - \log\det\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\right) - I, \tag{1.17}$$

where

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn} = \mathbf{x}_{fn}\mathbf{x}_{fn}^H \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\mathbf{x},fn} = \mathbf{R}_{jfn}v_{jfn}. \tag{1.18}$$

Note that the cost (1.17) is not well defined (i.e., its value is infinite) when $I > 1$ and matrices $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}$ are not full rank, which is the case in definition (1.18). However, this is not a problem per se. Indeed, the infinite term $-\log\det\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}\right)$ is independent on $\boldsymbol{\theta}$ and can be simply removed from the cost (1.17), since it has no influence on the optimization over $\boldsymbol{\theta}$. Otherwise, a small regularization term may be added to $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}$, which would make it full rank. Also, there exist alternative definitions of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}$ [42, 8], where it might be full rank by construction.

Formulation with the cost (1.17) is interesting, since, as one can note, it is a generalization of the IS-NMF cost in the single channel case (see Chapter *Févotte et al.*). Indeed, $C_{\mathrm{IS}}(\boldsymbol{\theta})$ becomes the single channel IS divergence when $I = 1$.

1.6.2 Maximum a posteriori

When a prior distribution $p(\boldsymbol{\theta})$ on model parameters is specified, like for example the spatial covariance prior in equation (1.13), the maximum a posteriori (MAP) criterion is usually used instead of the ML criterion. It writes

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}'} p(\boldsymbol{\theta}'|\mathbf{X}) = \arg\max_{\boldsymbol{\theta}'} p(\mathbf{X}|\boldsymbol{\theta}')p(\boldsymbol{\theta}'). \tag{1.19}$$

Note that in case of prior in (1.13) we have $p(\boldsymbol{\theta}) = \prod_{f=1}^{F} p\left(\mathbf{R}_{jf}|\boldsymbol{\Psi}_{jf}, m\right)^N$, since the prior is applied to each time-frequency bin.

If one tries rewriting (1.19) in a form similar to (1.17), it would result in simply adding $-\log p(\boldsymbol{\theta}')$ term to (1.17).

1.6.3 Other criteria

Several other criteria were proposed as well. For example, we have seen that the ML criterion formulated as in (1.17) generalizes the single channel IS NMF to the multichannel case, as such it was proposed in [16] to generalize the single-channel NMF with Euclidean distance (EUC NMF) to the multichannel case. This is achieved by replacing the cost function (1.17) with the following one

$$C_{\text{FRB}}(\boldsymbol{\theta}) = \sum_{f,n=1}^{F,N} \left\| \widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn} - \boldsymbol{\Sigma}_{\mathbf{x},fn} \right\|_F^2, \tag{1.20}$$

where $\|\mathbf{A}\|_F$ denotes the Frobenius norm of a matrix $\mathbf{A}$, and the data covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}$ is defined slightly differently than in (1.18). Notably, it is defined as [16, 17]

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn} = \sqrt{\left|\mathbf{x}_{fn}\mathbf{x}_{fn}^H\right|} \times \text{sign}\left(\mathbf{x}_{fn}\mathbf{x}_{fn}^H\right), \tag{1.21}$$

where all the operation, i.e., the absolute value $|\cdot|$, the square root $\sqrt{\cdot}$, the multiplication $\times$ and the sign ($\text{sign}(a) = a/|a|$), are applied element-wise to the corresponding matrices.

There is also the variational Bayes (VB) criterion [43], which consists in computing directly the posterior distribution of the source STFT coefficients while marginalizing over all possible model parameters.

1.7 Model estimation algorithms

There exist several model parameter estimation algorithms [8, 16]. Though, due to the probabilistic formulation of the LGM model (1.2), the expectation-maximization (EM) algorithm [44] is one of the most popular choices. As we will see below, the use of the EM algorithm results not in just one algorithm, but it leads to a family of algorithms. Indeed, each particular implementation of the EM algorithm depends on several choices, as will be explained below. Because of the EM popularity we

will mostly concentrate here on the different variants of EM and will only mention briefly other algorithms.

To present the variants of EM algorithm we consider the LGM model (1.2) with time-invariant unconstrained full rank spatial covariances $\mathbf{R}_{jf}$ and spatial variances v_{jfn} structured with NTF model (1.6). This is in fact a variant of multichannel NTF similar to the one described in [15], but with full rank covariances instead of rank-1 covariances as in [15]. Since no probabilistic priors on parameters are assumed, the variants of EM algorithm presented below are for the optimization of the ML criterion (1.16).

1.7.1 Variants of EM algorithm

In one of its general formulations the EM algorithm [44] to optimize the ML criterion (1.16) consists first in specifying

- so-called *observed data* $\mathbf{X}$ that are usually the multichannel mixture STFT coefficients in the case of multichannel source separation, as considered here, and
- so-called *latent data* $\mathbf{Z}$. The choice of latent data may be quite different and different choices would lead to different EM variants.

Assuming that a probabilistic model parametrized by $\boldsymbol{\theta}$ is specified, the EM algorithm is usually applied in the following case. It is applied when it is difficult to optimize in a closed form the ML criterion (1.16) maximizing $\log p(\mathbf{X}|\boldsymbol{\theta})$, while it is easy to maximize in a closed form or via some simplified iterative procedure the log-likelihood $\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$ of so-called *complete data* $\{\mathbf{X},\mathbf{Z}\}$. The choice of latent data $\mathbf{Z}$ is usually done accordingly.

The EM algorithm consists then in iterating the following two steps:

- **E-step**: Compute an auxiliary function as follows:

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(\ell)}) = \mathbb{E}_{\mathbf{X}|\mathbf{Z},\boldsymbol{\theta}^{(\ell)}} \log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}). \tag{1.22}$$

- **M-step**: Optimize the auxiliary function to update model parameters according to the following criterion:

$$\boldsymbol{\theta}^{(\ell+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(\ell)}), \tag{1.23}$$

  where $\boldsymbol{\theta}^{(\ell)}$ denotes the model parameters estimated at the ℓ-th iteration.

It is often possible to optimize the criterion (1.23) in a closed form. However, sometimes, depending on the choice of latent data $\mathbf{Z}$, it is not possible. In that case either another iterative optimization algorithm may be applied or any algorithm can be used provided that it assures at each iteration of EM the following non-decreasing of the auxiliary function:

$$Q(\boldsymbol{\theta}^{(\ell+1)}, \boldsymbol{\theta}^{(\ell)}) \geq Q(\boldsymbol{\theta}^{(\ell)}, \boldsymbol{\theta}^{(\ell)}). \tag{1.24}$$

In the latter case the algorithm is called generalized EM (GEM) [44], and the ways the optimization (1.24) is performed lead again to different variants of the algorithm.

To summarize let us list various choices that lead to different EM algorithm variants and thus different model parameters estimation results. These choices include:

1. Choice of latent data $\mathbf{Z}$, for example:
   - Latent data consist of NMF/NTF components [12] defined as

     $$c_{kjfn} \sim \mathcal{N}_c(0, w_{jfk} h_{jkn}), \qquad k = 1, \ldots, K_j \tag{1.25}$$

     in case of NMF spectral model (1.4), or as

     $$c_{kjfn} \sim \mathcal{N}_c(0, w_{fk} h_{kn} q_{jk}), \qquad k = 1, \ldots, K \tag{1.26}$$

     in case of NTF spectral model (1.6).
   - Latent data consist of so-called *sub-sources* [8] (see Sec. 1.7.2 below).
   - Latent data consist of point sources [15] s_{jfn} as in the narrowband approximation (1.8).
   - Latent data consist of spatial source images [27] $\mathbf{y}_{jfn}$ as in equation (1.2).
   - Latent data consist of binary TF activations of the predominant source (see, e.g., [45] for details).
2. Choice of maximization step updates in case of GEM algorithm, for example:
   - Closed-form updates in case of EM algorithm.
   - Alternating closed-form updates over subsets of parameters [27] (each subset of parameters is updated by a closed-form update, while the other parameters are fixed).
   - Multiplicative update (MU) rules [5] to update NMF/NTF spectral model parameters [8].
3. Choice of initial parameters $\boldsymbol{\theta}^{(0)}$, for example:
   - Random parameters initialization [8].
   - Parameters initialization using the source separation results obtained by a different algorithm [12].
4. Choice of number of EM algorithm iterations, for example:
   - Fixed number of iterations (the most common choice).
   - Iterating till some stopping criterion depending on the likelihood value is satisfied.

A so-called spatial image EM (SIEM) algorithm, where the latent data are the spatial source images, is given in details in the Chapter **Nugraha et al.**. In the following section we present in details a so-called sub-source EM algorithm based on MU rules (SSEM/MU) [8], where the latent data are the sub-sources and MU rules are used for the NTF spectral model parameters updates within the M-step. Other variants of the EM and GEM algorithms may be found in the corresponding papers.

1.7.2 Detailed presentation of SSEM/MU algorithm

Recall that our model consists of time-invariant unconstrained full rank spatial covariances $\mathbf{R}_{jf}$ and spatial variances v_{jfn} structured with NTF model (1.6). Thus, it can be parametrized as

$$\boldsymbol{\theta} = \left\{ \{\mathbf{R}_{jf}\}_{j,f}, \mathbf{Q}, \mathbf{W}, \mathbf{H} \right\}, \tag{1.27}$$

with nonnegative matrices $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ specified in Section 1.3.2.

The SSEM/MU algorithm presented below is a partial case of a more general algorithm from [8], though applied to a slightly different model (here the spectral variances are structured with NTF model, while in [8] they are structured with NMF model).

Each spatial $I \times I$ covariance $\mathbf{R}_{jf}$ being full rank, its rank equals to I. For each source j we introduce I so-called point *sub-sources* $s_{ji,fn} \in \mathbb{C}$ ($i = 1, \ldots, I$) that share the same spectral variance v_{jfn}, in other words they are distributed as

$$s_{ji,fn} \sim \mathcal{N}_c(0, v_{jfn}). \tag{1.28}$$

Moreover, each spatial covariance $\mathbf{R}_{jf}$ can be non-uniquely represented as

$$\mathbf{R}_{jf} = \mathbf{A}_{jf} \mathbf{A}_{jf}^H, \tag{1.29}$$

where $\mathbf{A}_{jf}$ is an $I \times I$ complex-valued matrix. By introducing a JI-length vector

$$\mathbf{s}_{fn} = \left[s_{11,fn}, \ldots, s_{1I,fn}, s_{21,fn}, \ldots, s_{2I,fn}, \cdots, s_{J1,fn}, \ldots, s_{JI,fn} \right]^T, \tag{1.30}$$

and an $I \times JI$ matrix

$$\mathbf{A}_f = \left[\mathbf{A}_{1f}, \mathbf{A}_{2f}, \ldots, \mathbf{A}_{Jf} \right], \tag{1.31}$$

one can show [8] that the LGM modeling (1.3) is equivalent (up to the noise term $\mathbf{b}_{fn}$) to

$$\mathbf{x}_{fn} = \mathbf{A}_{fn} \mathbf{s}_{fn} + \mathbf{b}_{fn}, \tag{1.32}$$

with $s_{ji,fn}$ (components of $\mathbf{s}_{fn}$) being mutually independent and distributed as in (1.28), the noise term $\mathbf{b}_{fn}$ being distributed as

$$\mathbf{b}_{fn} \sim \mathcal{N}_c(0, \boldsymbol{\Sigma}_{\mathbf{b},fn}), \tag{1.33}$$

with an anisotropic covariance matrix $\boldsymbol{\Sigma}_{\mathbf{b},fn} = \sigma^2_{\mathbf{b},f} \mathbf{I}_I$. The noise term $\mathbf{b}_{fn}$ is needed for a so-called *simulated annealing* procedure that is necessary in this case (see [12] for details), where the noise variance $\sigma^2_{\mathbf{b},f}$ is usually decreased over the algorithm iterations.

Let us now compute the auxiliary function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\ell)})$ defined in (1.22). Below we will omit sometimes the indexing of parameters with (ℓ), and it will be clear

from the context what are the parameters estimated on previous step and what are the parameters to be updated on the current step. The log-likelihood of the complete data $\{\mathbf{X},\mathbf{Z}\}$ writes [4]

$$\begin{aligned}\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) &= \log p(\mathbf{X}|\mathbf{Z},\boldsymbol{\theta}) + \log p(\mathbf{Z}|\boldsymbol{\theta}) \\ &\overset{\text{c}}{=} -\sum_{f,n} \mathbf{tr}\left[\boldsymbol{\Sigma}_{\mathbf{b},fn}^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{x},fn} - \mathbf{A}_{fn}\boldsymbol{\Sigma}_{\mathbf{xs},fn}^{H} - \boldsymbol{\Sigma}_{\mathbf{xs},fn}\mathbf{A}_{fn}^{H} + \mathbf{A}_{fn}\boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_{fn}^{H}\right)\right] \\ &\quad - \sum_{f,n}\log\left|\boldsymbol{\Sigma}_{\mathbf{b},fn}\right| - I\sum_{j,f,n} d_{IS}(\xi_{jfn}|v_{jfn}), \quad (1.34)\end{aligned}$$

where

$$\boldsymbol{\Sigma}_{\mathbf{x},fn} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn} = \mathbf{x}_{fn}\mathbf{x}_{fn}^{H} \quad (1.35)$$

is computed as in (1.18),

$$\boldsymbol{\Sigma}_{\mathbf{xs},fn} = \mathbf{x}_{fn}\mathbf{s}_{fn}^{H}, \quad (1.36)$$

$$\boldsymbol{\Sigma}_{\mathbf{s},fn} = \mathbf{s}_{fn}\mathbf{s}_{fn}^{H}, \quad (1.37)$$

$$\xi_{j,fn} = \frac{1}{I}\sum_{i=1}^{I}|s_{ji,fn}|^2, \quad (1.38)$$

and $d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$ is the scalar IS divergence (see Chapter *Févotte et al.*).

By applying the conditional expectation operator $\mathbb{E}_{\mathbf{X}|\mathbf{S},\boldsymbol{\theta}^{(\ell)}}[\cdot]$ the auxiliary function $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(\ell)})$ writes then

$$\begin{aligned}Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(\ell)}) &\overset{\text{c}}{=} -\sum_{f,n} \mathbf{tr}\left[\boldsymbol{\Sigma}_{\mathbf{b},fn}^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn} - \mathbf{A}_{fn}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xs},fn}^{H} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{xs},fn}\mathbf{A}_{fn}^{H} + \mathbf{A}_{fn}\widehat{\boldsymbol{\Sigma}}_{\mathbf{s},fn}\mathbf{A}_{fn}^{H}\right)\right] \\ &\quad - \sum_{f,n}\log\left|\boldsymbol{\Sigma}_{\mathbf{b},fn}\right| - I\sum_{j,f,n} d_{IS}(\hat{\xi}_{jfn}|v_{jfn}), \quad (1.39)\end{aligned}$$

with $\widehat{\boldsymbol{\Sigma}}_{\mathbf{xs},fn}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s},fn}$ and $\hat{\xi}_{jfn}$ defined as

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{xs},fn} = \mathbb{E}_{\mathbf{X}|\mathbf{S},\boldsymbol{\theta}^{(\ell)}}\left[\boldsymbol{\Sigma}_{\mathbf{xs},fn}\right], \quad (1.40)$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{s},fn} = \mathbb{E}_{\mathbf{X}|\mathbf{S},\boldsymbol{\theta}^{(\ell)}}\left[\boldsymbol{\Sigma}_{\mathbf{s},fn}\right], \quad (1.41)$$

$$\hat{\xi}_{jfn} = \mathbb{E}_{\mathbf{X}|\mathbf{S},\boldsymbol{\theta}^{(\ell)}}\left[\xi_{jfn}\right], \quad (1.42)$$

and computed as follows:

[4] When we write $\overset{\text{c}}{=}$, that means that the equality is up to some constant that is independent on model parameters $\boldsymbol{\theta}$, and thus has no influence on the optimization over parameters in (1.23).

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{xs},fn} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}\boldsymbol{\Omega}_{\mathbf{s},fn}^H, \tag{1.43}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{s},fn} = \boldsymbol{\Omega}_{\mathbf{s},fn}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}\boldsymbol{\Omega}_{\mathbf{s},fn}^H + (\mathbf{I}_{JI} - \boldsymbol{\Omega}_{\mathbf{s},fn}\mathbf{A}_f)\boldsymbol{\Sigma}_{\mathbf{s},fn}, \tag{1.44}$$

$$\hat{\xi}_{jfn} = \frac{1}{I}\sum_{i=(j-1)I+1}^{jI} \widehat{\boldsymbol{\Sigma}}_{\mathbf{s},fn}(i,i), \tag{1.45}$$

where

$$\boldsymbol{\Omega}_{\mathbf{s},fn} = \boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}, \tag{1.46}$$

$$\boldsymbol{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},fn}, \tag{1.47}$$

$$\boldsymbol{\Sigma}_{\mathbf{s},fn} = \operatorname{diag}\left(\left[\underbrace{v_{1,fn},\ldots,v_{1,fn}}_{I \text{ times}},\underbrace{v_{2,fn},\ldots,v_{2,fn}}_{I \text{ times}},\cdots,\underbrace{v_{J,fn},\ldots,v_{J,fn}}_{I \text{ times}}\right]\right). \tag{1.48}$$

We now proceed with the M-step (1.23). Maximizing the auxiliary function (1.39) over $\mathbf{A}_f$ leads to the following closed-form solution [5]:

$$\mathbf{A}_f = \widehat{\boldsymbol{\Sigma}}_{\mathbf{xs},fn}\widehat{\boldsymbol{\Sigma}}_{\mathbf{s},fn}^{-1}. \tag{1.49}$$

Maximization of the auxiliary function (1.39) over $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$, i.e., the minimization of $\sum_{j,f,n} d_{IS}(\hat{\xi}_{jfn}|v_{jfn})$ with v_{jfn} computed as in (1.6), does not allow a closed-form solution. As such, to update $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$, several iterations of the following MU rules [15] are applied:

$$q_{jk} \leftarrow q_{jk}\left(\frac{\sum_{f,n} w_{fk}h_{kn}\hat{\xi}_{jfn}v_{jfn}^{-2}}{\sum_{f,n} w_{fk}h_{kn}v_{jfn}^{-1}}\right), \tag{1.50}$$

$$w_{fk} \leftarrow w_{fk}\left(\frac{\sum_{j,n} h_{kn}q_{jk}\hat{\xi}_{jfn}v_{jfn}^{-2}}{\sum_{j,n} h_{kn}q_{jk}v_{jfn}^{-1}}\right), \tag{1.51}$$

$$h_{kn} \leftarrow h_{kn}\left(\frac{\sum_{j,f} w_{fk}q_{jk}\hat{\xi}_{jfn}v_{jfn}^{-2}}{\sum_{j,f} w_{fk}q_{jk}v_{jfn}^{-1}}\right). \tag{1.52}$$

Applying these MU rules does not guarantee auxiliary function minimization as in (1.23), but only its non-decreasing as in (1.24). As such, this is in fact a GEM algorithm.

Algorithm 1 summarizes one iteration of the SSEM/MU algorithm derived above.

[5] Note that if the spatial covariances $\mathbf{R}_{jf}$ are needed, they can be always computed with (1.29).

Algorithm 1 One iteration of SSEM/MU algorithm

- **E-step**: Compute statistics $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},fn}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{xs},fn}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s},fn}$ and $\hat{\xi}_{jfn}$ as in (1.35), (1.40), (1.41) and (1.42).
- **M-step**:
  - Update $\mathbf{A}_f$ as in (1.49).
  - Update $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ iterating equations (1.50), (1.51) and (1.52) several times.
  - Renormalize $\mathbf{A}_f$, $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ to remove scale ambiguity (see [12]).

1.7.3 Other algorithms

Another very popular choice for multichannel NMF model parameters estimation is the majorization-minimization (MM) algorithm [46], which is used for example in [16, 17]. Note that the EM algorithm is interpretable as a partial case of the MM algorithm.

1.8 Conclusion

In this chapter we have introduced multichannel NMF methods for audio source separation. Potential advantages and disadvantages of these methods are discussed. Despite a quickly growing popularity of deep learning that is now of a great interest for audio source separation, multichannel NMF methods remain still an important area of research and in our opinion cannot be completely replaced by deep learning-based methods in all situations. Indeed, especially in fully blind settings, where no training data are available, deep learning is not a suitable path any more, while multichannel NMF is still applicable.

As for the further research on multichannel NMF we would like highlighting the following possible paths which have been already started to be explored. One research direction consists in proposing more sophisticated spatial and spectral models adapted to the mixing conditions and sources of interest, as well as in proposing new models going beyond the limitations of the LGM modeling. Another direction consists in combining some aspects of multichannel NMF with deep learning.

Acknowledgment

Cédric Févotte acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 681839 (project FACTORY).

References

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

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

[3] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*, 2006.

[4] L. Le Magoarou, A. Ozerov, and N. Q. Duong, "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.

[5] 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.

[6] D. El Badawy, N. Q. Duong, and A. Ozerov, "On-the-fly audio source separationa novel user-friendly framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 261–272, 2017.

[7] 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.

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

[9] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[10] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. Irish Signals and Systems Conference*, Dublin, Ireland, Sep. 2005.

[11] ——, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, vol. 2008, no. Article ID 872425, p. 15 pages, 2008.

[12] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[13] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[14] M. I. Mandel, D. P. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *NIPS*, vol. 19, 2006.

[15] 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.

[16] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[17] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[18] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[19] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency gaussian source models," in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on.* IEEE, 2005, pp. 78–81.

[20] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local gaussian modeling," in *International Conference on Independent Component Analysis and Signal Separation.* Springer, 2009, pp. 775–782.

[21] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux, and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in *International Conference on Latent Variable Analysis and Signal Separation.* Springer, 2010, pp. 245–253.

[22] T. Higuchi, H. Takeda, T. Nakamura, and H. Kameoka, "A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden markov models." in *INTERSPEECH*, 2014, pp. 850–854.

[23] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1305–1322, 2005.

[24] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[25] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," in *International Conference on Independent Component Analysis and Signal Separation.* Springer, 2004, pp. 327–334.

[26] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[27] S. Arberet, A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE, 2010, pp. 1–4.

[28] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in models for acoustic processing, neural information processing systems workshop*. Citeseer, 2006.

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

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

[31] R. Bro, "Parafac. tutorial and applications," *Chemometrics and intelligent laboratory systems*, vol. 38, no. 2, pp. 149–171, 1997.

[32] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[33] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[34] N. Q. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, p. 149, 2013.

[35] R. Badeau and M. D. Plumbley, "Multichannel high-resolution nmf for modeling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 11, pp. 1670–1680, 2014.

[36] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "An inverse-gamma source variance prior with factorized parameterization for audio source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 136–140.

[37] N. Q. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 205–208.

[38] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," in

Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 3191–3195.

[39] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1408–1423, 2016.

[40] M. Togami, "Online speech source separation based on maximum likelihood of local gaussian modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 213–216.

[41] L. S. Simon and E. Vincent, "A general framework for online audio source separation," in *International conference on Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 397–404.

[42] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 73–80.

[43] K. Adiloğlu and E. Vincent, "Variational bayesian inference for source separation and robust feature extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1746–1758, 2016.

[44] A. P. Dempster, N. M. Laird, and D. B. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 39, pp. 1–38, 1977.

[45] J. Thiemann and E. Vincent, "A fast EM algorithm for Gaussian model-based source separation," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013, pp. 1–5.

[46] D. R. Hunter and K. Lange, "A tutorial on mm algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.