Gaussian model based multichannel separation

Alexey Ozerov, Hirokazu Kameoka

1
Gaussian model based multichannel separation

Alexey Ozerov and Hirokazu Kameoka

The Gaussian framework for multichannel source separation consists of modeling vectors of STFT coefficients as multivariate complex Gaussian distributions. It allows specifying spatial and spectral models of the source spatial images and estimating their parameters in a joint manner. *Multichannel nonnegative matrix factorization* illustrated in Fig. 1.1 is one of the most popular such methods. It combines nonnegative matrix factorization (NMF) (see Chapter **??**) and narrowband spatial modeling (see Chapter **??**). Besides NMF, the Gaussian framework makes it possible to reuse many other single-channel spectral models in a multichannel scenario. It differs from the frameworks in Chapters **??**, **??**, **??** by the fact that more advanced generative spectral models are typically used. Also, according to the general taxonomies introduced in Chapter **??**, it covers a wide range of audio source separation scenarios, including over- or underdetermined mixtures and weakly or strongly guided separation, and a wide range of methods, which are either learning-free or based on unsupervised/supervised source modeling.

In Section 1.1, we introduce the multichannel Gaussian framework. In Section 1.2, we provide a detailed list of spectral and spatial models. In Section, we explain how to estimate the parameters of these models 1.3. We give a detailed presentation of a few methods in Section 1.4 and provide a summary in Section 1.5.

1.1
Gaussian modeling

1.1.1
Joint spectral-spatial local Gaussian modeling

Let us start with the assumption that the narrowband approximation holds. Then, the $I \times 1$ spatial image $\mathbf{c}_j(n,f)$ of source j in time frame n and frequency bin f is modeled as the product of the acoustic transfer function $\mathbf{a}_j(f)$ and the short-time Fourier transform (STFT) coefficient $s_j(n,f)$ of source j:

$$\mathbf{c}_j(n,f) = \mathbf{a}_j(f)s_j(n,f). \tag{1.1}$$

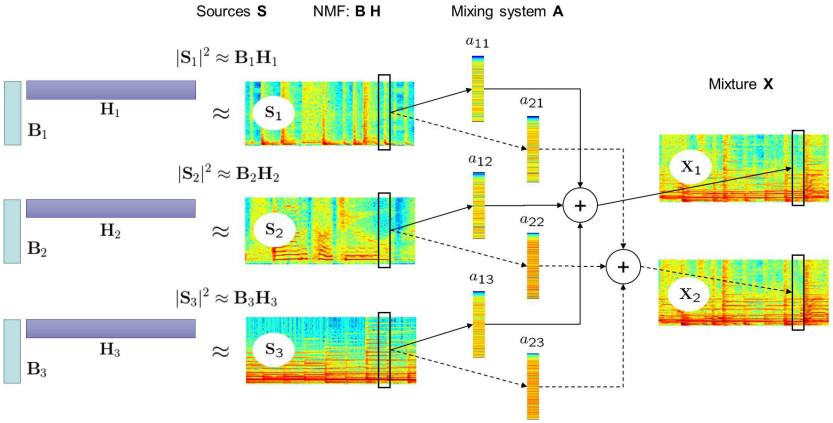

Figure 1.1 Illustration of multichannel NMF. $\mathbf{S}_j$, $\mathbf{X}_i$, and $\mathbf{a}_{ij}$ represent the complex-valued spectrograms of the sources and the mixture channels, and the complex-valued mixing coefficients, respectively. NMF factors the power spectrogram $|\mathbf{S}_j|^2$ of each source as $\mathbf{B}_j\mathbf{H}_j$ (see Chapter **??** and Section 1.2.1 below). The mixing system is represented by a rank-1 spatial model (see Section 1.2.2).

When $s_j(n, f)$ is assumed to follow a zero-mean complex Gaussian distribution with variance $\sigma_j^2(n, f)$

$$s_j(n, f) \sim \mathcal{N}_{\mathrm{c}}(s_j(n, f) \mid 0, \sigma_j^2(n, f)), \tag{1.2}$$

$\mathbf{c}_j(n, f)$ follows the so-called *local Gaussian model*

$$\mathbf{c}_j(n, f) \sim \mathcal{N}_{\mathrm{c}}(\mathbf{c}_j(n, f) \mid \mathbf{0}_I, \sigma_j^2(n, f)\mathbf{R}_j(f)) \tag{1.3}$$

where $\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^H(f)$ is the $I \times I$ spatial covariance matrix of source j. The narrowband approximation implies that the spatial covariance matrix has rank 1. Alternatively, $\mathbf{R}_j(f)$ can be assumed to be a full-rank matrix in (1.3). The local Gaussian model can also be defined using quadratic time-frequency representations instead of the STFT (Duong *et al.*, 2010b; Ozerov *et al.*, 2012).

Multichannel source separation problems can be formulated using this model (Pham *et al.*, 2003; Févotte and Cardoso, 2005; Vincent *et al.*, 2009; Duong *et al.*, 2010a; Sawada *et al.*, 2013; Higuchi and Kameoka, 2015). Let us show an example. The $I \times 1$ vector $\mathbf{x}(n, f)$ of STFT coefficients of the mixture signal is equal to the sum of the source spatial image vectors $\mathbf{c}_j(n, f)$ of J sources

$$\mathbf{x}(n, f) = \sum_{j=1}^{J} \mathbf{c}_j(n, f). \tag{1.4}$$

When the sources are assumed to be independent, $\mathbf{x}(n, f)$ follows

$$\mathbf{x}(n, f) \sim \mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n, f) \,\middle|\, \mathbf{0}_I, \sum_{j=1}^{J} \sigma_j^2(n, f)\mathbf{R}_j(f)\right). \tag{1.5}$$

Hence, we obtain the log-likelihood

$$\mathcal{M}^{\mathrm{ML}}(\boldsymbol{\theta}) = \sum_{nf} \left[-\log\det\left(\pi \sum_{j=1}^{J} \sigma_j^2(n,f)\mathbf{R}_j(f) \right) \right. \\ \left. -\mathbf{x}^H(n,f) \left(\sum_{j=1}^{J} \sigma_j^2(n,f)\mathbf{R}_j(f) \right)^{-1} \mathbf{x}(n,f) \right] \quad (1.6)$$

where $\boldsymbol{\theta} = \{\{\sigma_j^2(n,f)\}_{jnf}, \{\mathbf{R}_j(f)\}_{jf}\}$ is the set of unknown model parameters and $\mathcal{X} = \{\mathbf{x}(n,f)\}_{nf}$ the set of observed STFT coefficients. In the particular case when the narrowband approximation holds and there are as many sources as channels, i.e., $J = I$, the mixture (1.4) can be expressed as

$$\mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{s}(n,f) = (\mathbf{W}^H(f))^{-1}\mathbf{s}(n,f). \quad (1.7)$$

where $\mathbf{s}(n,f) = [s_1(n,f),\dots,s_I(n,f)]^T$ is the $I \times 1$ vector of source STFT coefficients, $\mathbf{A}(f) = [\mathbf{a}_1(f),\dots,\mathbf{a}_I(f)]$ is the $I \times I$ mixing matrix, and $\mathbf{W}^H(f) = \mathbf{A}^{-1}(f)$ is the $I \times I$ separation matrix. Hence (1.5) can be rewritten as

$$\mathbf{x}(n,f) \sim \mathcal{N}_c(\mathbf{x}(n,f) \mid \mathbf{0}_I, (\mathbf{W}^H(f))^{-1}\boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{W}^{-1}(f)), \quad (1.8)$$

with

$$\boldsymbol{\Sigma}_{\mathbf{s}}(n,f) = \mathrm{Diag}(\sigma_1^2(n,f),\dots,\sigma_I^2(n,f)). \quad (1.9)$$

This results in the log-likelihood of frequency-domain independent component analysis (FD-ICA) based on a time-varying Gaussian source model:

$$\mathcal{M}^{\mathrm{ML}}(\boldsymbol{\theta}) = \sum_{nf} \left[-I\log\pi + 2\log\det(\mathbf{W}(f)) - \sum_{j=1}^{I} \log\sigma_j^2(n,f) \right. \\ \left. - \mathbf{x}^H(n,f)\mathbf{W}(f)\boldsymbol{\Sigma}_{\mathbf{s}}^{-1}(n,f)\mathbf{W}^H(f)\mathbf{x}(n,f) \right]. \quad (1.10)$$

Since all the variables are indexed by frequency f in the log-likelihood, the optimization problem can be split into frequency-wise source separation problems. The permutation problem (see Section **??**) must then be solved in order to align the separated components in different frequency bins that originate from the same source. While some methods are designed to perform frequency binwise source separation followed by permutation alignment, it is preferable to solve permutation alignment and source separation in a joint manner since the clues used for permutation alignment can also be helpful for source separation.

To handle more general cases, such as when the sources outnumber the channels, or to solve the permutation and separation problems in a joint fashion, we must add further constraints to the local Gaussian model. In the following sections, we introduce assumptions and constraints that can be incorporated into the Gaussian framework in order to deal with various scenarios and to improve the source separation accuracy.

1.1.2 Source separation: main steps

Multichannel source separation methods based on the local Gaussian model can be categorized according to the choices of mixing models, source spectral models, spatial models, parameter estimation schemes and source signal estimation schemes. Here, we present the main steps to formulate these methods.

1.1.2.1 Mixing models

Typical choices of mixing models include additive, narrowband, subband filtering or sparse models. The three former models assume that all sources are active, while the sparse model assumes that only one source is active in each time-frequency bin. The additive model (1.4) makes no assumption about the source spatial images $\mathbf{c}_j(n,f)$, except that their sum is equal to the mixture $\mathbf{x}(n,f)$. The three latter models assume that there are J_p point sources of interest indexed by $j \in \{1,\dots,J_p\}$ and consider the other sources as background noise $\mathbf{u}(n,f) = \sum_{j=J_p+1}^{J} \mathbf{c}_j(n,f)$, which we assume to follow a zero-mean complex Gaussian distribution. The relationship between $\mathbf{x}(n,f)$, $\mathbf{s}(n,f) = [s_1(n,f),\dots,s_{J_p}(n,f)]^T$ and $\mathbf{u}(n,f)$ is defined for the narrowband mixing model by

$$\mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{s}(n,f) + \mathbf{u}(n,f) \tag{1.11}$$

$$= \sum_{j=1}^{J_p} \mathbf{a}_j(f)s_j(n,f) + \mathbf{u}(n,f), \tag{1.12}$$

for the subband filtering mixing model by

$$\mathbf{x}(n,f) = \sum_{n'=0}^{N'-1} \mathbf{A}(n',f)\mathbf{s}(n-n',f) + \mathbf{u}(n,f) \tag{1.13}$$

$$= \sum_{j=1}^{J_p}\sum_{n'=0}^{N'-1} \mathbf{a}_j(n',f)s_j(n-n',f) + \mathbf{u}(n,f), \tag{1.14}$$

and for the *sparse mixing model* by

$$\mathbf{x}(n,f) = \mathbf{a}_{z(n,f)}(f)s_{z(n,f)}(n,f) + \mathbf{u}(n,f), \tag{1.15}$$

where $z(n,f)$ denotes the index of the predominant source, i.e., the most active source in time-frequency bin (n,f). The length N' of the subband filters is typically in the order of L/M with L the length of the time-domain mixing filters $\mathbf{a}_j(\tau)$ and M the hop size between adjacent STFT frames.

In the particular case of a determined, noiseless mixture ($I = J$ and $\mathbf{u}(n,f) = \mathbf{0}_I$), the narrowband and the subband filtering mixing models can be inverted. Hence we can alternatively consider the narrowband demixing model

$$\mathbf{s}(n,f) = \mathbf{W}^H(f)\mathbf{x}(n,f) \tag{1.16}$$

or the truncated[1] subband filtering demixing model

$$\mathbf{s}(n,f) = \sum_{n'=0}^{N'-1} \mathbf{W}^H(n',f)\mathbf{x}(n-n',f). \tag{1.17}$$

1.1.2.2 **Source spectral models**

We assume that the STFT coefficients of source j follow (1.2). With this model, it can be shown using a simple change of variables that the power and phase of $s_j(n,f)$ follow an exponential distribution with mean $\sigma_j^2(n,f)$ and a uniform distribution on the interval $[0, 2\pi)$, respectively. If there is a certain assumption, constraint or structure that we want to impose on the power spectrum of each source, we can employ a parametric model to represent $\sigma_j^2(n,f)$ instead of individually treating $\sigma_j^2(n,f)$ as a free parameter, or introduce a properly designed prior distribution over $\sigma_j^2(n,f)$. Choices include a Gaussian mixture model (GMM) (Attias, 2003), a hidden Markov model (HMM) (Higuchi and Kameoka, 2015), an autoregressive (AR) model (Dégerine and Zaïdi, 2004; Yoshioka *et al.*, 2011), a nonnegative matrix/tensor factorization (NMF) model (Ozerov and Févotte, 2010; Arberet *et al.*, 2010; Ozerov *et al.*, 2011; Sawada *et al.*, 2013; Nikunen and Virtanen, 2014; Kitamura *et al.*, 2015), an excitation-filter model (also known as source-filter model) (Kameoka *et al.*, 2010; Ozerov *et al.*, 2012), a spectral continuity prior (Duong *et al.*, 2011), a deep neural network (DNN) model (Nugraha *et al.*, 2016), and combinations of different models (Ozerov *et al.*, 2012; Adiloğlu and Vincent, 2016), among others. These models will be presented in detail in Section 1.2.

1.1.2.3 **Spatial models**

The probability distribution of the observed signals $\mathcal{X} = \{\mathbf{x}(n,f)\}_{nf}$, i.e., the likelihood of the unknown parameters, can be derived according to the mixing model and the source distribution. For example, we can show from (1.11) and (1.2) that a narrowband mixture follows

$$\mathbf{x}(n,f) \sim \mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f) \;\middle|\; \mathbf{0}_I, \sum_{j=1}^{J_p} \sigma_j^2(n,f)\mathbf{R}_j(f) + \mathbf{\Sigma}_{\mathbf{u}}(f)\right) \tag{1.18}$$

where $\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^H(f)$ denotes the spatial covariance of source j and $\mathbf{\Sigma}_{\mathbf{u}}(f)$ is the noise covariance matrix. We can also show that a sparse mixture follows

$$\mathbf{x}(n,f) \mid z(n,f) \sim \mathcal{N}_{\mathrm{c}}(\mathbf{x}(n,f) \mid \mathbf{0}_I, \sigma_{z(n,f)}^2(n,f)\mathbf{R}_{z(n,f)}(f) + \mathbf{\Sigma}_{\mathbf{u}}(f)). \tag{1.19}$$

As with the source power spectrum, there are several ways to model the spatial covariance $\mathbf{R}_j(f)$. These models will be presented in detail in Section 1.2.2.

[1] The inverse of a finite impulse response (FIR) subband filter is generally an infinite impulse response filter.

1.1.2.4 **Parameter estimation schemes**

Let $\boldsymbol{\theta}$ be the set of parameters of the spectral and spatial models. Once the likelihood (and the prior distribution) of $\boldsymbol{\theta}$ has been defined according to the choice of mixing, spectral and spatial models, the next step is to derive a parameter estimation algorithm. Probabilistic parameter estimation schemes may be primarily divided into maximum likelihood (ML) or maximum a posteriori (MAP) estimation and Bayesian inference. The aim of the former is to find the estimate of $\boldsymbol{\theta}$ that maximizes the likelihood or the posterior distribution of $\boldsymbol{\theta}$ whereas the aim of the latter is to infer the posterior distribution of $\boldsymbol{\theta}$ given the observation $\mathcal{X}$. The typical choices of criteria and algorithms for parameter estimation will be presented in detail in Section 1.3.

1.1.2.5 **Source signal estimation schemes**

Once the parameters $\boldsymbol{\theta}$ have been estimated, we can estimate the source signals or their spatial images according to the assumed mixing model. In the case of the narrowband mixing model, a typical choice is the minimum mean square error (MMSE) estimator of $\mathbf{s}(n,f)$ (see Section **??**):

$$\widehat{\mathbf{s}}(n,f) = \mathbb{E}\{\mathbf{s}(n,f) \mid \mathbf{x}(n,f)\} = \mathbf{W}^H(n,f)\mathbf{x}(n,f), \tag{1.20}$$

where $\mathbf{W}(n,f)$ is the well-known multichannel Wiener filter (MWF):

$$\boldsymbol{\Sigma}_{\mathbf{s}}(n,f) = \mathrm{Diag}(\sigma_1^2(n,f),\ldots,\sigma_{J_p}^2(n,f)) \tag{1.21}$$

$$\mathbf{W}(n,f) = (\mathbf{A}(f)\boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{A}^H(f) + \boldsymbol{\Sigma}_{\mathbf{u}}(f))^{-1}\mathbf{A}(f)\boldsymbol{\Sigma}_{\mathbf{s}}(n,f). \tag{1.22}$$

When using a full-rank spatial covariance model, it may be convenient to use the MMSE estimator of the souce spatial image $\mathbf{c}_j(n,f)$ instead (see Section **??**):

$$\widehat{\mathbf{c}}_j(n,f) = \mathbb{E}\{\mathbf{c}_j(n,f) \mid \mathbf{x}(n,f)\} \tag{1.23}$$

$$= \sigma_j^2(n,f)\mathbf{R}_j(f)\left(\sum_{j'=1}^{J_p}\sigma_{j'}^2(n,f)\mathbf{R}_{j'}(f) + \boldsymbol{\Sigma}_{\mathbf{u}}(f)\right)^{-1}\mathbf{x}(n,f). \tag{1.24}$$

In the case of the narrowband and subband filtering demixing systems, we can directly use (1.16) and (1.17) (Dégerine and Zaïdi, 2004; Kameoka *et al.*, 2010; Yoshioka *et al.*, 2011; Kitamura *et al.*, 2015) once we have obtained the demixing filters $\mathbf{W}^H(f)$ or $\mathbf{W}^H(n',f)$. Algorithms for estimating the demixing filters are described in Section 1.4.3.

Finally, in the case of the sparse mixing system, one reasonable estimator is (Izumi *et al.*, 2007; Kameoka *et al.*, 2012)

$$\widehat{s}_j(n,f) = \gamma_j(n,f)\frac{\mathbf{a}_j^H(f)\boldsymbol{\Sigma}_{\mathbf{u}}^{-1}(n,f)\mathbf{x}(n,f)}{\mathbf{a}_j^H(f)\boldsymbol{\Sigma}_{\mathbf{u}}^{-1}(n,f)\mathbf{a}_j(f)} \tag{1.25}$$

which combines the source presence probability $\gamma_j(n,f) = P(z(n,f) = j \mid \mathcal{X}, \boldsymbol{\theta})$ and the minimum variance distortionless response (MVDR) beamformer.

Figure 1.2 schematizes the entire process via a block diagram.

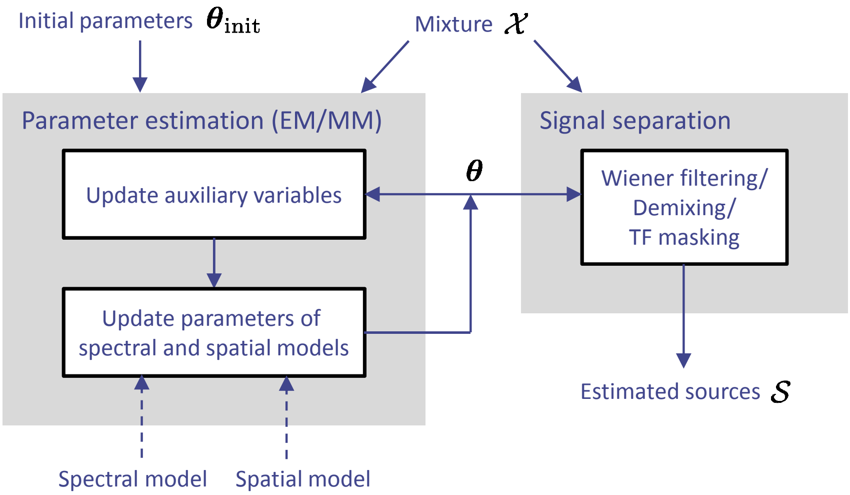
Initial parameters $\boldsymbol{\theta}_{\text{init}}$
Mixture $\mathcal{X}$
Parameter estimation (EM/MM)
Update auxiliary variables
Update parameters of spectral and spatial models
$\boldsymbol{\theta}$
Signal separation
Wiener filtering/ Demixing/ TF masking
Estimated sources $\mathcal{S}$
Spectral model
Spatial model

Figure 1.2 Block diagram of multichannel Gaussian model based source separation.

1.2 Library of spectral and spatial models

As mentioned in Section 1.1, the Gaussian framework reduces to FD-ICA in the case of determined noiseless mixtures if no specific constraint or structure is assumed on the power spectra and the spatial covariances of the sources. The advantages of incorporating source spectral models and spatial models into the Gaussian framework are twofold. First, they can help solving frequency-wise separation and permutation alignment in a joint fashion since the spectral and spatial properties represented by the models are useful for permutation alignment[2)]. Second, they allow us to deal with a larger range of mixtures, such as reverberant and/or underdetermined mixtures, by exploiting additional reasonable assumptions besides the independence of the sources. In this section, we present examples of spectral and spatial models.

1.2.1 Spectral models

1.2.1.1 GMM, scaled GMM, HMM

In speech, the number of phonemes and the pitch range are both usually limited during an entire utterance. In Western music, each piece of music is often played by only a handful of musical instruments or sung by one or a few singers and the number of musical notes is usually limited. It is thus reasonable to assume that the spectra of a real-world sound source can be described using a limited number of templates. By

[2)] Indeed, permutation alignment methods typically exploit the fact that the frequency components originating from the same source emanate from the same direction and that their magnitudes are correlated.

writing the spectral templates of source j as $b_{j,1}(f), \dots, b_{j,K_j}(f)$, where K_j denotes the number of spectral templates assigned to source j, one way to express the power spectrogram $\sigma_j^2(n,f)$ would be

$$\sigma_j^2(n,f) = b_{j,k_j(n)}(f), \tag{1.26}$$

where $k_j(n)$ denotes the index of the spectral template selected at frame n. If we assume $k_j(n)$ to be a latent variable generated according to a categorical distribution with probabilities $\pi_{j,1}, \dots, \pi_{j,K_j}$ such that $\sum_k \pi_{jk} = 1$, the generative process of the spatial image $\mathbf{c}_j(n,f)$ of source j is described as a GMM (Attias, 2003)[3]:

$$\mathbf{c}_j(n,f) \mid k_j(n) \sim \mathcal{N}_\mathrm{c}(\mathbf{c}_j(n,f) \mid \mathbf{0}_I, b_{j,k_j(n)}(f)\mathbf{R}_j(f)), \tag{1.27}$$

$$k_j(n) \sim \pi_{j,k_j(n)}. \tag{1.28}$$

Note that the spectral templates can be either trained on isolated signals of that source type in an unsupervised or a supervised manner or estimated from the mixture signal in a learning-free manner.

While the above model uses each template to represent a different power spectrum, it would be more reasonable to let each template represent all the power spectra that are equal up to a scale factor and treat the the scale factor as an additional parameter. Here, we use $b_{jk}(f)$ as the k-th "normalized" spectral template and describe $\sigma_j^2(n,f)$ as

$$\sigma_j^2(n,f) = b_{j,k_j(n)}(f)h_j(n), \tag{1.29}$$

where $h_j(n)$ denotes the time-varying amplitude. Note that this *scaled GMM* model has been employed by Benaroya *et al.* (2006) for single-channel source separation.

Furthermore, since the probability of a particular template being selected may depend on the templates selected at the previous frames, it is natural to extend the generative process of $k_j(n)$ using a Markov chain. These two extensions lead to a HMM (Vincent and Rodet, 2004; Ozerov *et al.*, 2009, 2012; Higuchi and Kameoka, 2015), namely

$$\mathbf{c}_j(n,f) \mid k_j(n) \sim \mathcal{N}_\mathrm{c}(\mathbf{c}_j(n,f) \mid \mathbf{0}_I, b_{j,k_j(n)}(f)h_j(n)\mathbf{R}_j(f)), \tag{1.30}$$

$$k_j(n) \mid k_j(n-1) \sim \pi_{j,k_j(n-1),k_j(n)}, \tag{1.31}$$

where (1.30) can be seen as the state emission probability, $k_j(n)$ as the hidden state, and $\pi_{jkk'}$ as the state transition probability from state k to state k' (see Fig. 1.3 top). By properly designing the state transition network, we can flexibly assign probabilities to state durations (the durations of the self-transitions). In addition, by incorporating states associated with speech absence or silence into the state transition network, assuming a state-dependent generative process of the scale factor $h_j(n)$ as

$$h_j(n) \mid k_j(n) \sim \mathcal{G}(h_j(n) \mid \alpha_{k_j(n)}, \beta_{k_j(n)}), \tag{1.32}$$

[3] Note that the use of GMM as a spectral model in multichannel Gaussian model based separation differs from its typical use in single-channel separation: in Section **??**, the GMM is nonzero-mean and it represents the distribution of the log-power spectrum, while here the GMM is zero-mean and it represents the distribution of the complex-valued STFT coefficients.

where $\mathcal{G}(\cdot \mid \alpha, \beta)$ denotes the gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$

$$\mathcal{G}(h \mid \alpha, \beta) = \frac{h^{\alpha-1}e^{-h/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \tag{1.33}$$

and setting the hyperparameters α_k and β_k so that $h_j(n)$ tends to be near zero for the states associated with speech absence, this model makes it possible to estimate voice activity segments along with solving the separation problem (Higuchi and Kameoka, 2015).

1.2.1.2 **NMF, NTF**

While the above models assume that only one of the spectral templates is activated at a time, another way to model the power spectrogram $\sigma_j^2(n, f)$ is to express it via NMF as the sum of the spectral templates $b_{j,1}(f), \dots, b_{j,K_j}(f)$ scaled by time-varying amplitudes $h_{j,1}(n), \dots, h_{j,K_j}(n)$ (see Fig. 1.3 bottom):

$$\sigma_j^2(n, f) = \sum_{k=1}^{K_j} b_{jk}(f)h_{jk}(n). \tag{1.34}$$

(1.34) can be interpreted as expressing the matrix $\widehat{\mathbf{V}}_j = [\sigma_j^2(n, f)]_{fn}$ as a product of two matrices $\mathbf{B}_j = [b_{jk}(f)]_{fk}$ and $\mathbf{H}_j = [h_{jk}(n)]_{kn}$. This leads to generative models of the STFT coefficients $s_j(n, f)$ (Févotte *et al.*, 2009) and the spatial image $\mathbf{c}_j(n, f)$ (Ozerov and Févotte, 2010)

$$s_j(n, f) \sim \mathcal{N}_{\mathrm{c}}\left(s_j(n, f) \,\middle|\, 0, \sum_k b_{jk}(f)h_{jk}(n)\right), \tag{1.35}$$

$$\mathbf{c}_j(n, f) \sim \mathcal{N}_{\mathrm{c}}\left(\mathbf{c}_j(n, f) \,\middle|\, \mathbf{0}_I, \sum_k b_{jk}(f)h_{jk}(n)\mathbf{R}_j(f)\right). \tag{1.36}$$

Multichannel source separation methods using this model or its variants are called multichannel NMF (Ozerov and Févotte, 2010; Kameoka *et al.*, 2010; Sawada *et al.*, 2013; Nikunen and Virtanen, 2014; Kitamura *et al.*, 2015). They generalize the single-channel Itakura-Saito (IS) NMF methods reviewed in Chapters **??** and **??** to the multichannel case.

With this model, the entire set of spectral templates is partitioned into subsets associated with individual sources. It is also possible to allow all the spectral templates to be shared by every source and let the contribution of the k-th spectral template to source j be determined in a learning-free manner (Ozerov *et al.*, 2011; Sawada *et al.*, 2013; Nikunen and Virtanen, 2014; Kitamura *et al.*, 2015). To do so, we drop the index j from $b_{jk}(f)$ and $h_{jk}(n)$, and instead introduce a continuous indicator variable $\phi_{jk} \geq 0$ such that $\sum_j \phi_{jk} = 1$. ϕ_{jk} can be interpreted as the expectation of a binary indicator variable, describing to which of the J sources the k-th template is assigned. The power spectrogram $\sigma_j^2(n, f)$ of source j can thus alternatively be

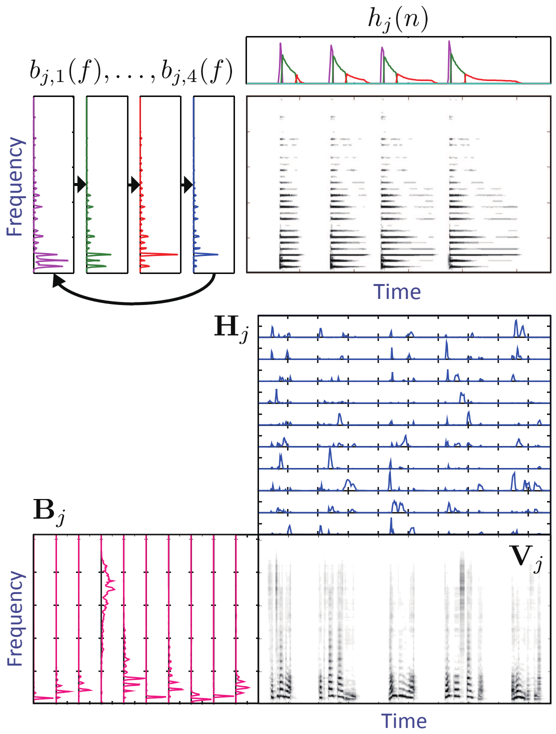

Figure 1.3 Illustration of the HMM and multichannel NMF spectral models.

modeled as

$$\sigma_j^2(n,f) = \sum_{k=1}^{K} \phi_{jk} b_k(f) h_k(n). \tag{1.37}$$

This spectral model is a form of nonnegative tensor factorization (NTF), which results in multichannel NTF.

1.2.1.3 AR and variants

Another reasonable assumption we can make about source power spectrograms is spectral continuity. This amounts to the assumption that the magnitudes of the STFT coefficients in all frequency bands originating from the same source tend to vary coherently over time. The most naive way would be to assume a flat spectrum with a time-varying scale

$$\sigma_j^2(n,f) = h_j(n). \tag{1.38}$$

This is actually a particular case of the NMF model (1.34) where $K_j = 1$ and $b_{j,1}(f) = 1$, which means each source has only one flat-shaped template. Under this constraint, assuming (1.35) is equivalent to assuming that the ℓ_2 norm

$\|[s_j(n,0),\dots,s_j(n,F-1)]^T\|_2 = \sqrt{\sum_f |s_j(n,f)|^2}$ follows a Gaussian distribution with time-varing variance $h_j(n)$. This is analogous to the assumption employed by independent vector analysis (IVA) (see Section **??**) where the ℓ_2 norm is assumed to follow a supergaussian distribution, which is shown to be effective in eliminating the inherent permutation indeterminacy of FD-ICA.

Other representations ensuring spectral continuity include the AR model (also known as the all-pole model) (Dégerine and Zaïdi, 2004; Yoshioka *et al.*, 2011)

$$\sigma_j^2(n,f) = \frac{\sigma_j^2(n)}{|1-\alpha_1(n)e^{-2\jmath\pi f/F} - \cdots - \alpha_{N'}(n)e^{-2\jmath\pi N'f/F}|^2}, \tag{1.39}$$

where $\alpha_1(n),\dots,\alpha_{N'}(n)$ denote the AR parameters at time n and N' is the number of poles. This expression is justified by the fact that the power spectrum of speech can be approximated fairly well by an excitation-filter representation using an all-pole model as the vocal tract filter. A combination of the AR model and the NMF model has also been proposed (Kameoka and Kashino, 2009; Kameoka *et al.*, 2010). With this model, the power spectrum of a source is expressed as the sum of all possible pairs of excitation and filter templates scaled by time-varying amplitudes

$$\sigma_j^2(n,f) = \sum_k \sum_{l=1}^{L} \frac{b_{jk}(f)h_{jkl}(n)}{|1-\alpha_{jl,1}e^{-2\jmath\pi f/F} - \cdots - \alpha_{jl,N'}e^{-2\jmath\pi N'f/F}|^2}, \tag{1.40}$$

where $b_{jk}(f)$ denotes the k-th excitation spectral template, the denominator is the l-th all-pole vocal tract spectral template, and $h_{jkl}(n)$ denotes the time-varying amplitude of the (k,l)-th excitation-filter pair of source j. We can easily confirm that when $L=1$ and $N'=0$, this model reduces to the NMF model (1.34). Note that these spectral templates can be either pretrained using training samples or estimated from the mixture signal in a learning-free manner.

Another way to impose a certain structure on $\sigma_j^2(n,f)$ is to place a prior distribution over $\sigma_j^2(n,f)$. For example, a prior distribution for ensuring spectral continuity can be designed using an inverse-gamma chain (Duong *et al.*, 2011)

$$\sigma_j^2(n,f) \mid \sigma_j^2(n,f-1) \sim \mathcal{IG}(\sigma_j^2(n,f) \mid \alpha, (\alpha-1)\sigma_j^2(n,f-1)), \tag{1.41}$$

where $\mathcal{IG}(\cdot \mid \alpha,\beta)$ denotes the inverse gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$

$$\mathcal{IG}(v \mid \alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} v^{-\alpha-1} e^{-\beta/v}, \tag{1.42}$$

whose mean is $\beta/(\alpha-1)$. It is also possible to ensure temporal continuity by assuming

$$\sigma_j^2(n,f) \mid \sigma_j^2(n-1,f) \sim \mathcal{IG}(\sigma_j^2(n,f) \mid \alpha, (\alpha-1)\sigma_j^2(n-1,f)). \tag{1.43}$$

These priors can be combined with NMF (see Section **??**).

1.2.1.4 **Composite models and DNN**

A general flexible framework with various combinations of these spectral models is presented by Ozerov *et al.* (2012) and Adiloğlu and Vincent (2016). It should also be noted that a DNN-based approach has been proposed recently (Nugraha *et al.*, 2016), where DNNs are used to model the source power spectrograms and combined with the Gaussian framework to exploit the spatial information (refer to Section **??** for details).

1.2.2 Spatial models

Spatial modeling consists of constraining the spatial covariances $\mathbf{R}_j(f)$ in (1.3). Constraints are usually introduced by reparameterizing $\mathbf{R}_j(f)$, by imposing some prior distribution on it, or both.

Assuming that the narrowband approximation (1.11) holds, the spatial covariance may simply be constrained as

$$\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^H(f), \tag{1.44}$$

which restricts the rank of $\mathbf{R}_j(f)$ to be 1. This model is called the rank-1 model. It was used by Févotte and Cardoso (2005) and Ozerov and Févotte (2010) and many other authors. Alternatively, it was later proposed by Duong *et al.* (2010a) and Sawada *et al.* (2013) to consider an unconstrained full-rank model for $\mathbf{R}_j(f)$. This model partly overcomes the limitations of the narrowband approximation, and it better handles mixtures with long reverberation times (RT60s).

A popular way of constraining spatial models is to introduce constraints related to the source direction of arrival (DOA). Izumi *et al.* (2007) simply constrained the rank-1 model to $\mathbf{R}_j(f) = \tilde{\mathbf{d}}(\alpha_j, f)\tilde{\mathbf{d}}^H(\alpha_j, f)$, with $\tilde{\mathbf{d}}(\alpha_j, f)$ the relative steering vector (**??**) corresponding to the j-th source DOA α_j. The unknown source DOAs α_j are then inferred from the mixture. Duong *et al.* (2010a) extended this expression to the full-rank model by adding the covariance matrix of the diffuse reverberation field (see Section **??**). Duong *et al.* (2013) allowed some deviation from this constraint by setting an inverse-Wishart prior on $\mathbf{R}_j(f)$ whose mean is parameterized by the DOA.

Alternatively, Nikunen and Virtanen (2014) consider a DOA grid $\{\alpha_k\}_{k=1}^K$ and constrain the spatial covariance matrix of each source as

$$\mathbf{R}_j(f) = \sum_{k=1}^{K} q_{jk}\tilde{\mathbf{d}}(\alpha_k, f)\tilde{\mathbf{d}}^H(\alpha_k, f), \tag{1.45}$$

with nonnegative weights q_{jk}. Such a combination of DOA-based rank-1 models (also called DOA kernels) makes it possible to model not only the direct path, but also reflections. Kameoka *et al.* (2012) and Higuchi and Kameoka (2015) proposed similar *DOA mixture models* where the acoustic transfer function $\mathbf{a}_j(f)$ within the rank-1 model (Kameoka *et al.*, 2012) or the full-rank model $\mathbf{R}_j(f)$ (Higuchi and

Kameoka, 2015) are distributed as K-component mixture models and the distribution of each component constrains the corresponding spatial model to be close to a predefined DOA α_k.

1.3 Parameter estimation criteria and algorithms

1.3.1 Parameter estimation criteria

Once a spectral model and a spatial model have been specified for each source, a criterion must be chosen for model parameter estimation. Let $\boldsymbol{\theta}$ denote the full set of parameters of the chosen models. For example, in the case of NMF spectral models and full-rank spatial models (Arberet *et al.*, 2010) $\boldsymbol{\theta}$ consists of the NMF parameters and the full-rank spatial covariance matrices of all sources. Specifying a parameter estimation criterion resides in defining a cost or an objective function to be optimized over $\boldsymbol{\theta}$ given a multichannel mixture $\mathcal{X}$.

ML is one of the most popular criteria (Ozerov and Févotte, 2010; Duong *et al.*, 2010a; Sawada *et al.*, 2013). It consists of maximizing the log-likelihood

$$\mathcal{M}^{\mathrm{ML}}(\boldsymbol{\theta}) = \log p(\mathcal{X} \mid \boldsymbol{\theta}). \tag{1.46}$$

In case of the local Gaussian model (1.3), this is equivalent to minimizing the cost

$$\mathcal{C}^{\mathrm{IS}}(\boldsymbol{\theta}) = \sum_{nf} \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f)) - \log\det(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f)) - I, \tag{1.47}$$

where $\boldsymbol{\Sigma}_{\mathbf{x}}(n,f) = \sum_j \sigma_j^2(n,f)\mathbf{R}_j(f)$ is the model covariance and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f) = \mathbb{E}\{\mathbf{x}(n,f)\mathbf{x}^H(n,f)\}$ is an estimate of the data covariance (Ozerov *et al.*, 2012)[4]. This cost is a multichannel extension of the IS divergence (see Section **??**).

We see that optimizing criterion (1.47) consists in minimizing a measure of fit between the model covariance $\boldsymbol{\Sigma}_{\mathbf{x}}(n,f)$ and the data covariance $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)$. Sawada *et al.* (2013) and Nikunen and Virtanen (2014) proposed to replace this measure of fit by the Frobenius norm. This leads to the cost

$$\mathcal{C}^{\mathrm{EUC}}(\boldsymbol{\theta}) = \sum_{nf} \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f) - \boldsymbol{\Sigma}_{\mathbf{x}}(n,f)\|_F^2 \tag{1.48}$$

that is a multichannel generalization of the squared Euclidean (EUC) distance (see Section **??**). The computation of the model and data covariances is then modified such that they scale with the magnitude of the data, since the EUC distance is usually applied to magnitude spectra rather than power spectra in the single-channel case.

[4] Contrary to (1.46), (1.47) takes finite values only when $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)$ is full-rank. To cope with that, the term $-\log\det\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)$ may be removed from the cost, since it is independent from $\boldsymbol{\theta}$.

MAP estimation is an alternative to ML (1.46) that maximizes the log-posterior

$$\mathcal{M}^{\mathrm{MAP}}(\boldsymbol{\theta}) = \log p(\mathcal{X}, \boldsymbol{\theta}) = \log p(\mathcal{X} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \tag{1.49}$$

with a suitable prior distribution $p(\boldsymbol{\theta})$ on the model parameters. Using MAP instead of ML results in the additional term $-\log p(\boldsymbol{\theta})$ in the corresponding cost functions (1.47) or (1.48).

As opposed to ML (1.46) and MAP (1.49), where a point estimate of the model parameters $\boldsymbol{\theta}$ is sought, *variational Bayesian* (VB) inference (Kameoka *et al.*, 2012; Adiloğlu and Vincent, 2016; Kounades-Bastian *et al.*, 2016) aims to estimate the posterior distribution of the source spatial images $\mathcal{C} = \{\mathbf{c}_j(n, f)\}_{jnf}$ while marginalizing over all possible model parameters:

$$p(\mathcal{C} \mid \mathcal{X}) = \frac{p(\mathcal{C}, \mathcal{X})}{p(\mathcal{X})} = \frac{\int p(\mathcal{C}, \boldsymbol{\theta}, \mathcal{X}) \mathrm{d}\boldsymbol{\theta}}{\iint p(\mathcal{C}, \boldsymbol{\theta}, \mathcal{X}) \mathrm{d}\mathcal{C} \mathrm{d}\boldsymbol{\theta}}. \tag{1.50}$$

Since the integrals in (1.50) are computationally intractable, a factored approximation of the joint posterior $p(\mathcal{C}, \boldsymbol{\theta}, \mathcal{X})$ is assumed. The criterion to be minimized is then the Kullback-Leibler (KL) divergence between the true posterior and the factored approximation (see Section 1.3.2.3 below for details).

1.3.2 Parameter estimation algorithms

1.3.2.1 EM algorithm

We here formulate the *expectation-maximization* (EM) algorithm (Dempster *et al.*, 1977) as applied to optimize the MAP criterion (1.49), since it is more general than the ML criterion (1.46) and it reduces to ML in the case of a noninformative prior $p(\boldsymbol{\theta}) \propto 1$. In most cases $\mathcal{M}^{\mathrm{MAP}}(\boldsymbol{\theta})$ has several local and global maxima, and there is no closed-form solution for a global maximum.

To find a local maximum the EM algorithm consists of first defining so-called *latent data* (also called *hidden data*) $\mathcal{Z}$ and then iterating the following two steps:

- *E-step*: Compute the posterior distribution of the latent data $p(\mathcal{Z} \mid \mathcal{X}, \boldsymbol{\theta}^{(m)})$ and derive the auxiliary function[5]

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathcal{Z}|\mathcal{X},\boldsymbol{\theta}^{(m)}} \left\{ \log \frac{p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta})}{p(\mathcal{Z} \mid \mathcal{X}, \boldsymbol{\theta}^{(m)})} \right\}. \tag{1.51}$$

- *M-step*: Update the model parameter estimates so as to maximize the auxiliary function:

$$\boldsymbol{\theta}^{(m+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}), \tag{1.52}$$

where $\boldsymbol{\theta}^{(m)}$ are the model parameter estimates obtained in the m-th iteration.

[5] The term $-\mathbb{E}_{\mathcal{Z}}\{\log p(\mathcal{Z} \mid \mathcal{X}, \boldsymbol{\theta}^{(m)})\}$ does not depend on $\boldsymbol{\theta}$ hence it is often omitted in the expression of the $\mathcal{Q}$ function.

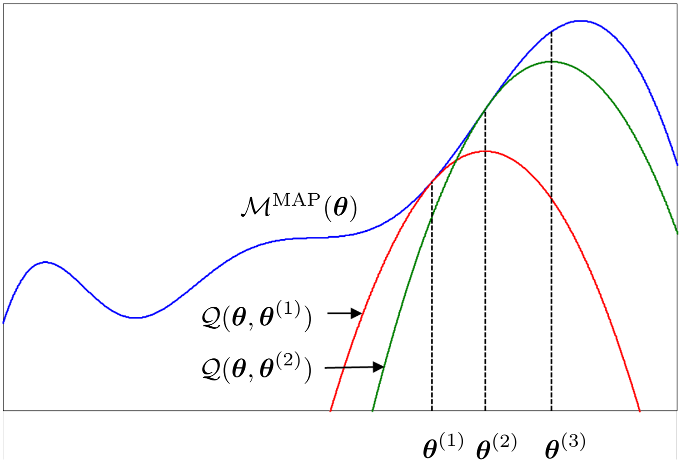
$\mathcal{M}^{\text{MAP}}(\boldsymbol{\theta})$
$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(1)})$
$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(2)})$
$\boldsymbol{\theta}^{(1)}$
$\boldsymbol{\theta}^{(2)}$
$\boldsymbol{\theta}^{(3)}$

Figure 1.4 Graphical illustration of the EM algorithm for MAP estimation.

It can be shown that $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \leq \mathcal{M}^{\text{MAP}}(\boldsymbol{\theta})$ and $\mathcal{Q}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) = \mathcal{M}^{\text{MAP}}(\boldsymbol{\theta}^{(m)})$, i.e., the auxiliary function is a lower bound of the log-posterior that is tight at the current solution $\boldsymbol{\theta}^{(m)}$. With this property, it can be proved that each iteration of the above EM algorithm does not decrease the value of $\mathcal{M}^{\text{MAP}}(\boldsymbol{\theta})$ (Dempster *et al.*, 1977). This can be intuitively understood from the graphical illustration in Fig. 1.4. A relaxed variant of the EM algorithm called generalized EM consists in replacing the M-step's closed-form maximization of the auxiliary function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ by any update that makes it nondecreasing, i.e., $\mathcal{Q}(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\theta}^{(m)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$.

It is worth to note that even for the same ML or MAP criterion there may be various ways of implementing the EM algorithm. Indeed, each implementation, and as a consequence the final result, depends on the choice of the latent data $\mathcal{Z}$, the M-step parameter update in the case of the generalized EM algorithm, and the initial parameter values $\boldsymbol{\theta}^{(0)}$.

In the case of multichannel NMF, Ozerov and Févotte (2010) and Arberet *et al.* (2010) define the NMF components $\mathbf{y}_{jk}(n, f)$ such that $\mathbf{c}_j(n, f) = \sum_{k=1}^{K_j} \mathbf{y}_{jk}(n, f)$ and each component $\mathbf{y}_{jk}(n, f)$ has a zero-mean Gaussian distribution with covariance $b_{jk}(f)h_{jk}(n)\mathbf{R}_j(f)$[6] and they consider these components as latent data. Alternatively, Duong *et al.* (2010a) consider directly the source images $\mathcal{C}$ as latent data. Ozerov *et al.* (2012) showed that a source image with a rank-r spatial model can be represented as the sum of r *subsources* each modeled by a rank-1 spatial model. Considering those subsources as latent data makes it possible to specify a unified EM algorithm suitable for spatial models of any rank (Ozerov *et al.*, 2012). In the case of the sparse mixing model (1.15), the indices $z(n, f)$ of the active sources are typically considered then as latent data instead, which allows consider-

[6] This model is equivalent to (1.36).

able computational savings in the resulting EM algorithm (Thiemann and Vincent, 2013).

Several approaches (Ozerov *et al.*, 2011, 2012) consider the source spatial images or subsources as latent data and employ the multiplicative update rules of single-channel IS-NMF (see Section **??**) within the M-step, which results in variants of the generalized EM algorithm. These approaches, which are usually referred to as (generalized) EM with multiplicative updates, often allow speeding up the algorithm's convergence (Ozerov *et al.*, 2011).

1.3.2.2 **MM algorithm**

The *majorization-minimization* (MM) algorithm (also known as *auxiliary function-based* optimization) (Leeuw and Heiser, 1977; Hunter and Lange, 2004) is a generalization of the EM algorithm. When constructing an MM algorithm for a given minimization problem, the main issue is to design an auxiliary function called a majorizer that is guaranteed to never go below the cost function. If such a majorizer is properly designed, an algorithm that iteratively minimizes the majorizer is guaranteed to converge to a stationary point of the cost function. The MM algorithm was used for single-channel NMF by several authors (Lee and Seung, 2000; Nakano *et al.*, 2010; Févotte and Idier, 2011). In general, if we can build a tight majorizer that is easy to optimize, we can expect to obtain a fast converging algorithm.

Suppose $\mathcal{C}(\boldsymbol{\theta})$ is a cost function that we want to minimize with respect to $\boldsymbol{\theta}$. A majorizer $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is defined as a function satisfying

$$\mathcal{C}(\boldsymbol{\theta}) = \min_{\boldsymbol{\alpha}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\alpha}) \tag{1.53}$$

where $\boldsymbol{\alpha}$ is an auxiliary variable. $\mathcal{C}(\boldsymbol{\theta})$ can then be shown to be nonincreasing under the updates

$$\boldsymbol{\theta} \leftarrow \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\alpha}) \tag{1.54}$$

$$\boldsymbol{\alpha} \leftarrow \operatorname*{argmin}_{\boldsymbol{\alpha}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\alpha}). \tag{1.55}$$

This can be proved as follows. Let us denote the iteration number by m, set $\boldsymbol{\theta}$ to an arbitrary value $\boldsymbol{\theta}^{(m)}$ and define $\boldsymbol{\alpha}^{(m+1)} = \operatorname{argmin}_{\boldsymbol{\alpha}} \mathcal{Q}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\alpha})$ and $\boldsymbol{\theta}^{(m+1)} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\alpha}^{(m+1)})$. First, it is obvious that $\mathcal{C}(\boldsymbol{\theta}^{(m)}) = \mathcal{Q}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\alpha}^{(m+1)})$. Next, we can confirm that $\mathcal{Q}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\alpha}^{(m+1)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\alpha}^{(m+1)})$ since $\boldsymbol{\theta}^{(m+1)}$ is the minimizer of $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\alpha}^{(m+1)})$ with respect to $\boldsymbol{\theta}$. By definition, it is obvious that $\mathcal{Q}(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\alpha}^{(m+1)}) \geq \mathcal{C}(\boldsymbol{\theta}^{(m+1)})$ and so we can finally show that $\mathcal{C}(\boldsymbol{\theta}^{(m)}) \geq \mathcal{C}(\boldsymbol{\theta}^{(m+1)})$.

Here, we briefly show that the EM algorithm is a special case of the MM algorithm. Let $\mathcal{X}$ be the observed data, $\mathcal{C}^{\text{MAP}}(\boldsymbol{\theta}) = -\log p(\mathcal{X}, \boldsymbol{\theta})$ the cost function that we want to minimize with respect to the parameters $\boldsymbol{\theta}$, and $\mathcal{Z}$ the latent data. The latent data can be either discrete or continuous. While we consider the continuous case here, the following also applies to the discrete case by simply replacing the integral over $\mathcal{Z}$ with a summation. First, we can show that

$$\mathcal{C}^{\mathrm{MAP}}(\boldsymbol{\theta}) = -\log \int p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta}) \mathrm{d}\mathcal{Z} \tag{1.56}$$

$$= -\log \int \lambda(\mathcal{Z}) \frac{p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta})}{\lambda(\mathcal{Z})} \mathrm{d}\mathcal{Z} \tag{1.57}$$

$$\leq -\int \lambda(\mathcal{Z}) \log \frac{p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta})}{\lambda(\mathcal{Z})} \mathrm{d}\mathcal{Z}, \tag{1.58}$$

where $\lambda(\mathcal{Z})$ is an arbitrary nonnegative weight function that is subject to the normalization constraint

$$\int \lambda(\mathcal{Z}) \mathrm{d}\mathcal{Z} = 1. \tag{1.59}$$

(1.58) follows from Jensen's inequality by using the fact that the negative logarithm is a convex function. We can use the right-hand side of this inequality as the majorizer of $\mathcal{C}^{\mathrm{MAP}}(\boldsymbol{\theta})$. Thus, we can show that $\mathcal{C}^{\mathrm{MAP}}(\boldsymbol{\theta})$ is nonincreasing under the updates

$$\lambda(\mathcal{Z}) \leftarrow \underset{\lambda(\mathcal{Z})}{\operatorname{argmin}} - \int \lambda(\mathcal{Z}) \log \frac{p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta})}{\lambda(\mathcal{Z})} \mathrm{d}\mathcal{Z} = p(\mathcal{Z} \mid \mathcal{X}, \boldsymbol{\theta}) \tag{1.60}$$

$$\boldsymbol{\theta} \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmin}} - \int \lambda(\mathcal{Z}) \log \frac{p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta})}{\lambda(\mathcal{Z})} \mathrm{d}\mathcal{Z}. \tag{1.61}$$

(1.60) stems from the fact that the inequality in (1.58) becomes an equality when

$$\frac{p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta})}{\lambda(\mathcal{Z})} = \xi(\mathcal{X}, \boldsymbol{\theta}), \tag{1.62}$$

is independent of $\mathcal{Z}$, which yields

$$\lambda(\mathcal{Z}) = \frac{p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta})}{\xi(\mathcal{X}, \boldsymbol{\theta})} \tag{1.63}$$

$$\Rightarrow \int \lambda(\mathcal{Z}) d\mathcal{Z} = \frac{1}{\xi(\mathcal{X}, \boldsymbol{\theta})} \int p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta}) \mathrm{d}\mathcal{Z} = 1 \tag{1.64}$$

$$\Rightarrow \xi(\mathcal{X}, \boldsymbol{\theta}) = \int p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta}) \mathrm{d}\mathcal{Z} = p(\mathcal{X}, \boldsymbol{\theta}) \tag{1.65}$$

$$\Rightarrow \lambda(\mathcal{Z}) = \frac{p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta})}{p(\mathcal{X}, \boldsymbol{\theta})} = p(\mathcal{Z} \mid \mathcal{X}, \boldsymbol{\theta}). \tag{1.66}$$

We can confirm that (1.60) and (1.61) correspond to the expectation and maximization steps, respectively.

1.3.2.3 **VB algorithm**

The VB approach is another extension of EM which aims to estimate the posterior distribution of all the random variables involved in the generative model. Let $\boldsymbol{\theta}$ be the entire set of variables of interest (including, e.g., the source STFT coefficients, the model parameters, and the latent data) and $\mathcal{X}$ be the observed data. Our goal is to compute the posterior

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\boldsymbol{\theta}, \mathcal{X})}{p(\mathcal{X})}. \tag{1.67}$$

The joint distribution $p(\boldsymbol{\theta}, \mathcal{X})$ can usually be written explicitly according to the assumed generative model. However, to obtain the exact posterior $p(\boldsymbol{\theta} \mid \mathcal{X})$, we must compute $p(\mathcal{X})$, which involves an intractable integral. Instead of obtaining the exact posterior, the VB approach considers approximating this posterior variationally by minimizing

$$\mathcal{C}^{\mathrm{VB}}(q(\boldsymbol{\theta})) = \mathcal{C}^{\mathrm{KL}}(q(\boldsymbol{\theta}) \mid p(\boldsymbol{\theta} \mid \mathcal{X})), \tag{1.68}$$

with respect to $q(\boldsymbol{\theta})$ with

$$\int q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 1, \tag{1.69}$$

where $\mathcal{C}^{\mathrm{KL}}(\cdot \mid \cdot)$ denotes the KL divergence

$$\mathcal{C}^{\mathrm{KL}}(q(\boldsymbol{\theta}) \mid p(\boldsymbol{\theta} \mid \mathcal{X})) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathcal{X})} \mathrm{d}\boldsymbol{\theta}. \tag{1.70}$$

By partitioning the set of variables as $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}_k$ and restricting the class of approximate distributions to those that factorize into

$$q(\boldsymbol{\theta}) = \prod_k q(\boldsymbol{\theta}_k) \quad \text{with} \quad \int q(\boldsymbol{\theta}_k) \mathrm{d}\boldsymbol{\theta}_k = 1, \tag{1.71}$$

we can use a simple block coordinate descent algorithm to find a local minimum of (1.68) for each factor in turn. It can be shown using the calculus of variations that the optimal distribution for each factor is

$$q(\boldsymbol{\theta}_k) \propto \exp[\mathbb{E}_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_k)}\{\log p(\boldsymbol{\theta}, \mathcal{X})\}], \tag{1.72}$$

where $\mathbb{E}_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_k)}\{\log p(\boldsymbol{\theta}, \mathcal{X})\}$ is the expectation of the joint probability of the data and the variables, taken over all variables except $\boldsymbol{\theta}_k$.

1.3.3 Categorization of existing methods

Table 1.1 categorizes various approaches discussed above according to the underlying mixing model, spectral model, spatial model, estimation criterion and algorithm.

1.4 Detailed presentation of some methods

We now give detailed descriptions of two popular parameter estimation algorithms. For both algorithms, we consider the narrowband mixing model (1.11), the full-rank unconstrained spatial model (see Section 1.2.2), and the NTF spectral model (1.37).

Method	Mixing	Spatial model	Spectral model	Criterion	Algorithm
Attias (2003)	subband filter mix	rank-1	GMM	VB	VB
Izumi *et al.* (2007)	sparse	rank-1	unconstrained	EUC	EM
Duong *et al.* (2010b)	additive	full-rank	unconstrained	IS	generalized EM
Kameoka *et al.* (2010)	subband filter demix	rank-1	NMF+AR	IS	EM
Ozerov and Févotte (2010)	narrowband mix	rank-1	NMF	IS	generalized EM
Yoshioka *et al.* (2011)	subband filter demix	rank-1	AR	IS	block coord. descent
Kameoka *et al.* (2012)	sparse	rank-1 DOA mixture	unconstrained	VB	VB
Ozerov *et al.* (2012)	additive	any rank	NMF/GMM/ excit.-filter/…	IS	generalized EM
Duong *et al.* (2013)	additive	full-rank DOA prior	unconstrained	IS (MAP)	generalized EM
Sawada *et al.* (2013)	additive	full-rank	NMF	IS/ EUC	MM
Nikunen and Virtanen (2014)	narrowband mix	rank-1 DOA kernels	NMF	EUC	MM
Higuchi and Kameoka (2015)	subband filter mix	full-rank DOA mixture	HMM	IS	MM
Kitamura *et al.* (2015)	subband filter demix	rank-1	NMF	IS	MM
Adiloğlu and Vincent (2016)	narrowband mix	rank-1	NMF/ excit.-filter	VB	VB
Nugraha *et al.* (2016)	additive	full-rank	DNN	IS	EM

Table 1.1 Categorization of existing approaches according to the underlying mixing model, spectral model, spatial model, estimation criterion and algorithm.

1.4.1 IS multichannel NTF EM algorithm

The EM algorithm presented below is a combination of those presented by Ozerov and Févotte (2010), Arberet *et al.* (2010), and Ozerov *et al.* (2011). More specifically the spatial full-rank model is that of Arberet *et al.* (2010), the spectral NTF model is that of Ozerov *et al.* (2011), and the choice of the NTF components as latent data follows Ozerov and Févotte (2010).

Let us introduce the NTF components $\mathbf{y}_{jk}(n,f)$ such that $\mathbf{c}_j(n,f) = \sum_{k=1}^{K_j} \mathbf{y}_{jk}(n,f)$ and each component $\mathbf{y}_{jk}(n,f)$ is distributed as

$$\mathbf{y}_{jk}(n,f) \sim \mathcal{N}_{\mathrm{c}}(\mathbf{y}_{jk}(n,f) \mid \mathbf{0}_I, \mathbf{\Sigma}_{\mathbf{y}_{jk}}(n,f)) \tag{1.73}$$

with

$$\mathbf{\Sigma}_{\mathbf{y}_{jk}}(n,f) = \phi_{jk} b_k(f) h_k(n) \mathbf{R}_j(f). \tag{1.74}$$

This formulation is strictly equivalent to the original model. We denote the full set of model parameters as $\boldsymbol{\theta} = \{\{\phi_{jk}\}_{jk}, \{b_k(f)\}_{kf}, \{h_k(n)\}_{kn}, \{\mathbf{R}_j(f)\}_{jf}\}$.

Following Ozerov and Févotte (2010), we consider the set of NTF components $\mathcal{Y} = \{\mathbf{y}_{jk}(n,f)\}_{jknf}$ as latent data. Assuming a noninformative prior $p(\boldsymbol{\theta}) \propto 1$, the auxiliary function (1.51) for the ML criterion can be written as

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) &= \mathbb{E}_{\mathcal{Y}|\mathcal{X},\boldsymbol{\theta}^{(m)}}\{\log p(\mathcal{Y} \mid \boldsymbol{\theta})\} + \mathrm{cst}(\boldsymbol{\theta}^{(m)}) \\ &= \sum_{jknf} -\log\det\left(\pi \boldsymbol{\Sigma}_{\mathbf{y}_{jk}}(n,f)\right) \\ &\qquad - \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jk}}(n,f)\boldsymbol{\Sigma}^{-1}_{\mathbf{y}_{jk}}(n,f)) + \mathrm{cst}(\boldsymbol{\theta}^{(m)}) \end{aligned} \tag{1.75}$$

where the term $\mathrm{cst}(\boldsymbol{\theta}^{(m)})$ depends only on $\boldsymbol{\theta}^{(m)}$ and is independent of $\boldsymbol{\theta}$, thus has no influence on the optimization in (1.52), and

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jk}}(n,f) &= \mathbb{E}_{\mathcal{Y}|\mathcal{X},\boldsymbol{\theta}^{(m)}}\{\mathbf{y}_{jk}(n,f)\mathbf{y}^H_{jk}(n,f)\} \\ &= \widehat{\mathbf{y}}_{jk}(n,f)\widehat{\mathbf{y}}^H_{jk}(n,f) + (\mathbf{I}_I - \mathbf{W}^H_{jk})\boldsymbol{\Sigma}^{(m)}_{\mathbf{y}_{jk}}(n,f)\end{aligned} \tag{1.76}$$

with $\mathbf{I}_I$ the $I \times I$ identity matrix and

$$\mathbf{W}_{jk}(n,f) = \left(\sum_{j'k'} \boldsymbol{\Sigma}^{(m)}_{\mathbf{y}_{j'k'}}(n,f)\right)^{-1} \boldsymbol{\Sigma}^{(m)}_{\mathbf{y}_{jk}}(n,f) \tag{1.77}$$

$$\widehat{\mathbf{y}}_{jk}(n,f) = \mathbf{W}^H_{jk}(n,f)\mathbf{x}(n,f). \tag{1.78}$$

Note that the maximum of $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ over $\boldsymbol{\theta}$ has no closed-form solution. However, it is possible to compute a closed-form maximum for each of the four parameter subsets given the other three subsets. Alternately maximizing each subset guarantees that the auxiliary function is nondecreasing. Thus, we obtain a generalized EM algorithm that can be summarized as follows:

- **E-step**: Compute the statistics $\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jk}}(n,f)$ as in (1.76).
- **M-step**: Update the model parameters $\boldsymbol{\theta}$ as

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_n \frac{1}{\sum_{jk}\phi_{jk}b_k(f)h_k(n)}\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jk}}(n,f) \tag{1.79}$$

$$\widehat{v}_{jk}(n,f) = \frac{1}{I}\,\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jk}}(n,f)) \tag{1.80}$$

$$\phi_{jk} = \frac{1}{NF}\sum_{nf}\frac{\widehat{v}_{jk}(n,f)}{b_k(f)h_k(n)} \tag{1.81}$$

$$b_k(f) = \frac{1}{JN}\sum_{jn}\frac{\widehat{v}_{jk}(n,f)}{\phi_{jk}h_k(n)} \tag{1.82}$$

$$h_k(n) = \frac{1}{JF}\sum_{jf}\frac{\widehat{v}_{jk}(n,f)}{\phi_{jk}b_k(f)}. \tag{1.83}$$

1.4.2
IS multichannel NMF MM algorithm

We now give a detailed description of the MM algorithm for multichannel NMF of Sawada *et al.* (2013). This algorithm is an extension of the MM algorithm originally developed by Kameoka *et al.* (2006) for solving general model fitting problems using the IS divergence. We show first how to derive the MM algorithm for single-channel NMF with the IS divergence and then how to extend it to the multichannel case.

The cost function for single-channel NMF with the IS divergence can be written as

$$\mathcal{C}^{\mathrm{IS}}(\boldsymbol{\theta}) = \sum_{nf} \left(\frac{|x(n,f)|^2}{\sigma^2(n,f)} + \log v(n,f) \right), \tag{1.84}$$

where $x(n,f)$ are the observed STFT coefficients, $\sigma^2(n,f) = \sum_k b_k(f)h_k(n)$ and $\boldsymbol{\theta}$ is a set consisting of $\mathbf{B} = [b_k(f)]_{kf}$ and $\mathbf{H} = [h_k(n)]_{kn}$ (Févotte *et al.*, 2009). Although it is difficult to obtain a closed-form expression of the global minimum, a majorizer of $\mathcal{C}^{\mathrm{IS}}(\boldsymbol{\theta})$ can be obtained as follows (Kameoka *et al.*, 2006). First, by using the fact that the function $f(x) = 1/x$ is convex for $x > 0$, we can use Jensen's inequality to obtain

$$\frac{|x(n,f)|^2}{\sigma^2(n,f)} \leq \sum_k \rho_k(n,f) \frac{|x(n,f)|^2}{b_k(f)h_k(n)/\rho_k(n,f)} = \sum_k \rho_k^2(n,f) \frac{|x(n,f)|^2}{b_k(f)h_k(n)}, \tag{1.85}$$

where $\rho_k(n,f) \geq 0$ is an arbitrary weight that must satisfy $\sum_k \rho_k(n,f) = 1$. It can be shown that the equality holds when

$$\rho_k(n,f) = \frac{b_k(f)h_k(n)}{\sum_{k'} b_{k'}(f)h_{k'}(n)}. \tag{1.86}$$

Next, since the function $f(x) = \log x$ is concave for $x > 0$, the tangent to $f(x)$ is guaranteed to never lie below $f(x)$. Thus, we have

$$\log \sigma^2(n,f) \leq \frac{\sigma^2(n,f) - \kappa(n,f)}{\kappa(n,f)} + \log \kappa(n,f) \tag{1.87}$$

for any $\kappa(n,f) > 0$. The equality holds when

$$\kappa(n,f) = \sigma^2(n,f). \tag{1.88}$$

By combining these inequalities, we have

$$\mathcal{C}^{\mathrm{IS}}(\boldsymbol{\theta}) \leq \sum_{nf} \Bigg(\sum_k \rho_k^2(n,f) \frac{|x(n,f)|^2}{b_k(f)h_k(n)} \\ + \frac{\sigma^2(n,f) - \kappa(n,f)}{\kappa(n,f)} + \log \kappa(n,f) \Bigg). \tag{1.89}$$

Hence, we can use the right-hand side of this inequality as a majorizer for $\mathcal{C}^{\mathrm{IS}}(\boldsymbol{\theta})$ where $\{\rho_k(n,f)\}_{knf}$ and $\{\kappa(n,f)\}_{nf}$ are auxiliary variables. Here, (1.86) and (1.88) correspond to the update rules for the auxiliary variables. What is particularly notable about this majorizer is that while $\mathcal{C}^{\mathrm{IS}}(\boldsymbol{\theta})$ involves nonlinear interaction of $b_1(f)h_1(n), \dots, b_K(f)h_K(n)$, it is given in a separable form expressed as a sum of the $1/b_k(f)h_k(n)$ and $b_k(f)h_k(n)$ terms, which are relatively easy to optimize with respect to $b_k(f)$ and $h_k(n)$. By differentiating this majorizer with respect to $b_k(f)$ and $h_k(n)$, and setting the results to zero, we obtain the following update rules for $b_k(f)$ and $h_k(n)$:

$$b_k(f) = \sqrt{\frac{\sum_n \rho_k^2(n,f)|x(n,f)|^2/h_k(n)}{\sum_n h_k(n)/\kappa(n,f)}} \tag{1.90}$$

$$h_k(n) = \sqrt{\frac{\sum_f \rho_k^2(n,f)|x(n,f)|^2/b_k(f)}{\sum_f b_k(f)/\kappa(n,f)}}. \tag{1.91}$$

Now, let us turn to the cost function (1.47) for multichannel NMF where

$$\boldsymbol{\Sigma}_{\mathbf{x}}(n,f) = \sum_{jk} \phi_{jk} b_k(f) h_k(n) \mathbf{R}_j(f). \tag{1.92}$$

We can confirm that when the number of channels and sources is $I = 1$ and $J = 1$, respectively, and $\phi_{jk} = 1$, this cost function reduces to the cost (1.84). We can obtain a majorizer given in a separable form in the same way as the single-channel case. By analogy with (1.85), we have

$$\begin{aligned} &\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f)) \\ &\qquad\qquad \le \sum_{jk} \frac{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{P}_{jk}(n,f)\mathbf{R}_j^{-1}(f)\mathbf{P}_{jk}(n,f))}{\phi_{jk} b_k(f) h_k(n)} \end{aligned} \tag{1.93}$$

for the first term with an arbitrary $I \times I$ complex-valued matrix $\mathbf{P}_{jk}(n,f)$ such that $\sum_{jk} \mathbf{P}_{jk}(n,f) = \mathbf{I}_I$, and

$$\begin{aligned} &\log\det(\boldsymbol{\Sigma}_{\mathbf{x}}(n,f)) \\ &\qquad\qquad \le \mathrm{tr}(\mathbf{K}^{-1}(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}(n,f)) + \log\det\mathbf{K}(n,f) - I \end{aligned} \tag{1.94}$$

for the second term with a positive definite matrix $\mathbf{K}(n,f)$ (Sawada *et al.*, 2013). We can show that the equalities in (1.93) and (1.94) hold when

$$\mathbf{P}_{jk}(n,f) = \phi_{jk} b_k(f) h_k(n) \mathbf{R}_j(f) \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f) \tag{1.95}$$

$$\mathbf{K}(n,f) = \boldsymbol{\Sigma}_{\mathbf{x}}(n,f). \tag{1.96}$$

By combining these inequalities, we have

$$\mathcal{C}^{\mathrm{IS}}(\boldsymbol{\theta}) \leq \sum_{nf} \Bigg[\sum_{jk} \frac{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{P}_{jk}(n,f)\mathbf{R}_j^{-1}(f)\mathbf{P}_{jk}(n,f))}{\phi_{jk} b_k(f) h_k(n)} + \mathrm{tr}(\mathbf{K}^{-1}(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}(n,f)) + \log\det \mathbf{K}(n,f) - I \Bigg]. \quad (1.97)$$

Hence, we can use the right-hand side of this inequality as a majorizer for $\mathcal{C}^{\mathrm{IS}}(\boldsymbol{\theta})$ where $\mathcal{P} = [\mathbf{P}_{jk}(n,f)]_{jknf}$ and $\mathcal{K} = [\mathbf{K}(n,f)]_{nf}$ are auxiliary variables. Here, (1.95) and (1.96) correspond to the update rules for the auxiliary variables. Similarly to the single-channel case, this majorizer is given in a separable form, which is relatively easy to optimize with respect to $\boldsymbol{\Phi} = [\phi_{jk}]_{jk}$, $\mathbf{B} = [b_k(f)]_{kf}$, $\mathbf{H} = [h_k(n)]_{kn}$ and $\mathcal{R} = [\mathbf{R}_j(f)]_{jf}$. By differentiating this majorizer with respect to $b_k(f)$ and $h_k(n)$ and setting the results to zero, we obtain the following update rules for $b_k(f)$ and $h_k(n)$:

$$b_k(f) = \sqrt{\frac{\displaystyle\sum_{jn} \frac{1}{\phi_{jk} h_k(n)} \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{P}_{jk}(n,f)\mathbf{R}_j^{-1}(f)\mathbf{P}_{jk}(n,f))}{\displaystyle\sum_{jn} \phi_{jk} h_k(n)\, \mathrm{tr}(\mathbf{K}^{-1}(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}(n,f))}} \quad (1.98)$$

$$h_k(n) = \sqrt{\frac{\displaystyle\sum_{jf} \frac{1}{\phi_{jk} b_k(f)} \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{P}_{jk}(n,f)\mathbf{R}_j^{-1}(f)\mathbf{P}_{jk}(n,f))}{\displaystyle\sum_{jf} \phi_{jk} b_k(f)\, \mathrm{tr}(\mathbf{K}^{-1}(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}(n,f))}}. \quad (1.99)$$

As regards ϕ_{jk}, although it is necessary to take the unit sum constraint into account, here we describe a convenient approach, which consists of updating ϕ_{jk} as

$$\phi_{jk} = \sqrt{\frac{\displaystyle\sum_{nf} \frac{1}{b_k(f) h_k(n)} \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{P}_{jk}(n,f)\mathbf{R}_j^{-1}(f)\mathbf{P}_{jk}(n,f))}{\displaystyle\sum_{nf} b_k(f) h_k(n)\, \mathrm{tr}(\mathbf{K}^{-1}(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}(n,f))}}, \quad (1.100)$$

which minimizes the majorizer, and projecting it onto the constraint space as $\phi_{jk} \leftarrow \phi_{jk}/\sum_{j'} \phi_{j',k}$, followed by rescaling of $b_k(f)$ and $h_k(n)$. As regards $\mathbf{R}_j(f)$, the optimal update is given as the solution of the algebraic Riccati equation

$$\mathbf{R}_j(f)\boldsymbol{\Psi}_j(f)\mathbf{R}_j(f) = \boldsymbol{\Omega}_j(f), \quad (1.101)$$

where the coefficient matrices are given by

$$\boldsymbol{\Psi}_j(f) = \sum_{kn} \phi_{jk} b_k(f) h_k(n) \mathbf{K}^{-1}(n,f) \tag{1.102}$$

$$\boldsymbol{\Omega}_j(f) = \sum_{kn} \frac{\mathbf{P}_{jk}(n,f)\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{P}_{jk}(n,f)}{\phi_{jk} b_k(f) h_k(n)}. \tag{1.103}$$

Since there is a scale indeterminacy between $\mathbf{R}_j(f)$ and $\phi_{jk} b_k(f) h_k(n)$, a convenient way to eliminate the indeterminacy is to update $\mathbf{R}_j(f)$ using the above equation and then perform unit trace normalization: $\mathbf{R}_j(f) \leftarrow \mathbf{R}_j(f)/\operatorname{tr}(\mathbf{R}_j(f))$.

Sawada *et al.* (2013) compare the convergence of the EM algorithm and the MM algorithm for IS multichannel NMF.

1.4.3 **Other algorithms for demixing filter estimation**

For the narrowband (1.16) and subband filtering (1.17) demixing models, one popular way for estimating the demixing filters $\mathbf{W}^H(f)$ involves the natural gradient method (Amari *et al.*, 1996). Here, we show other useful methods using block coordinate descent (Ono, 2011; Kameoka *et al.*, 2010; Yoshioka *et al.*, 2011).

First, let us consider the narrowband case (1.16). Recall that the log-likelihood of $\mathbf{W}(f) = [\mathbf{w}_1(f), \ldots, \mathbf{w}_J(f)]$ is given by (1.10). When the log-likelihood is given in this form, it can be maximized analytically with respect to one of the column vectors of $\mathbf{W}(f)$. Thus, we use a block coordinate descent algorithm to estimate $\mathbf{W}(f)$ by iteratively minimizing the negative log-likelihood with respect to each column vector while keeping the other column vectors fixed (Ono, 2011; Kitamura *et al.*, 2015). By keeping only the terms that depend on $\mathbf{W}(f)$ in the negative log-likelihood, the cost function for $\mathbf{W}(f)$ can be written as

$$\mathcal{C}^{\mathrm{ML}}(\mathbf{W}(f)) = N \sum_j \mathbf{w}_j^H(f) \boldsymbol{\Sigma}_{\mathbf{x}/\sigma_j}(f) \mathbf{w}_j(f) - 2N \log\det(\mathbf{W}(f)) + \mathrm{cst}, \tag{1.104}$$

where $\boldsymbol{\Sigma}_{\mathbf{x}/\sigma_j}(f) = \frac{1}{N}\sum_n \frac{\mathbf{x}(n,f)\mathbf{x}^H(n,f)}{\sigma_j^2(n,f)}$. By computing the complex derivative of $\mathcal{C}^{\mathrm{ML}}(\mathbf{W}(f))$ with respect to the conjugate of one column vector $\mathbf{w}_j^*(f)$[7], and setting the result to zero, we have

$$\boldsymbol{\Sigma}_{\mathbf{x}/\sigma_j}(f)\mathbf{w}_j(f) - 2\frac{\partial}{\partial \mathbf{w}_j^*(f)} \log\det(\mathbf{W}(f)) = \mathbf{0}_I. \tag{1.105}$$

By using the matrix formula $(\partial/\partial \mathbf{W}^*)\det(\mathbf{W}) = (\mathbf{W}^{-1})^H \det(\mathbf{W})$, (1.105) can be rearranged in the following simultaneous vector equations

$$\mathbf{w}_j^H(f)\boldsymbol{\Sigma}_{\mathbf{x}/\sigma_j}(f)\mathbf{w}_j(f) = 1 \tag{1.106}$$

$$\mathbf{w}_{j'}^H(f)\boldsymbol{\Sigma}_{\mathbf{x}/\sigma_j}(f)\mathbf{w}_j(f) = 0 \text{ for } j' \neq j. \tag{1.107}$$

[7] For complex-valued differentiation and matrix formulas, see Petersen and Pedersen (2005).

A solution to (1.106) and (1.107) can be found by the following updates:

$$\mathbf{w}_j(f) \leftarrow (\mathbf{W}^H(f)\mathbf{\Sigma}_{\mathbf{x}/\sigma_j}(f))^{-1}\mathbf{e}_j \tag{1.108}$$

$$\mathbf{w}_j(f) \leftarrow \frac{\mathbf{w}_j(f)}{\sqrt{\mathbf{w}_j^H(f)\mathbf{\Sigma}_{\mathbf{x}/\sigma_j}(f)\mathbf{w}_j(f)}}, \tag{1.109}$$

where $\mathbf{e}_j$ denotes the j-th column of the $J \times J$ identity matrix $\mathbf{I}_J$.

Next, let us turn to the subband filtering case (1.17). When $\mathbf{W}^H(0, f)$ is invertible, (1.17) can be written equivalently as the following process

$$\mathbf{y}(n, f) = \mathbf{x}(n, f) - \sum_{n'=1}^{N'-1} \tilde{\mathbf{W}}^H(n', f)\mathbf{x}(n - n', f), \tag{1.110}$$

$$\mathbf{s}(n, f) = \mathbf{W}^H(0, f)\mathbf{y}(n, f), \tag{1.111}$$

where $\tilde{\mathbf{W}}^H(n', f) = -(\mathbf{W}^H(0, f))^{-1}\mathbf{W}^H(n', f)$ (Yoshioka *et al.*, 2011; Kameoka *et al.*, 2010). (1.110) can be seen as a dereverberation process of the observed mixture signal $\mathbf{x}(n, f)$ described as a multichannel AR system with regression matrices $\tilde{\mathcal{W}} = \{\tilde{\mathbf{W}}^H(n', f)\}_{n'f}$ whereas (1.111) can be seen as a narrowband demixing process of the dereverberated mixture signal $\mathbf{y}(n, f)$. When $\mathbf{W}^H(0, f)$ is fixed, it can be shown that the log-likelihood of $\tilde{\mathcal{W}}$ becomes equal up to a sign to the objective function of a vector version of the linear prediction problem (also called multichannel linear prediction), which can be maximized with respect to $\tilde{\mathcal{W}}$ by solving a Yule-Walker equation. When $\tilde{\mathcal{W}}$ is fixed, on the other hand, the log-likelihood of $\mathbf{W}^H(0, f)$ becomes equal up to a sign and constant terms to (1.104) with $\mathbf{x}(n, f)$ replaced with $\mathbf{y}(n, f)$, namely $\mathbf{\Sigma}_{\mathbf{y}/\sigma_j}(f) = \frac{1}{N}\sum_n \mathbf{y}(n, f)\mathbf{y}^H(n, f)/\sigma_j^2(n, f)$, which can be locally maximized with respect to $\mathbf{W}^H(0, f)$ using the natural gradient method or the method described above. Thus, we can find estimates of $\mathbf{W}^H(0, f)$ and $\tilde{\mathcal{W}}$ by optimizing each of them in turn (Yoshioka *et al.*, 2011; Kameoka *et al.*, 2010).

1.5 Summary

The Gaussian framework for multichannel source separation is particularly noteworthy in that it provides a flexible way to incorporate source spectral models and spatial covariance models into a generative model of multichannel signals so that it can combine various clues to handle reverberation, underdetermined mixtures, and permutation alignment problems. It is also remarkable in that it allows us to develop powerful and efficient algorithms for parameter inference and estimation, taking advantage of the properties of Gaussian random variables. In this chapter, we presented the main steps to formulate Gaussian model-based methods, examples of source spectral models and spatial models along with the motivations behind them, and detailed derivations of several popular algorithms for multichannel NMF. For extensions to moving sources or microphones, refer to Chapter **??**.

Acknowledgment

We thank E. Vincent for help with writing this chapter.

Bibliography

Adiloğlu, K. and Vincent, E. (2016) Variational Bayesian inference for source separation and robust feature extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (10), 1746–1758.

Amari, S., Cichocki, A., and Yang, H.H. (1996) A new learning algorithm for blind signal separation, in *Proceedings of Neural Information Processing Systems*, pp. 757–763.

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

Attias, H. (2003) New EM algorithms for source separation and deconvolution with a microphone array, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, vol. V, vol. V, pp. 297–300.

Benaroya, L., Bimbot, F., and Gribonval, R. (2006) Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, **14** (1), 191–199.

Dégerine, S. and Zaïdi, A. (2004) Separation of an instantaneous mixture of Gaussian autoregressive sources by the exact maximum likelihood approach. *IEEE Transactions on Signal Processing*, **52** (6), 1499–1512.

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

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

Duong, N.Q.K., Vincent, E., and Gribonval, R. (2010a) Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, **18** (7), 1830–1840.

Duong, N.Q.K., Vincent, E., and Gribonval, R. (2010b) Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation, in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation*, pp. 73–80.

Duong, N.Q.K., Vincent, E., and Gribonval, R. (2013) Spatial location priors for Gaussian model based reverberant audio source separation. *EURASIP Journal on Advances in Signal Processing*, **2013**, 149.

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

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

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

Higuchi, T. and Kameoka, H. (2015) Unified approach for audio source separation with multichannel factorial HMM and DOA mixture model, in *Proceedings of European Signal Processing Conference*, pp. 2043–2047.

Hunter, D.R. and Lange, K. (2004) A tutorial on MM algorithms. *The American*

Statistician, **58** (1), 30–37.

Izumi, Y., Ono, N., and Sagayama, S. (2007) Sparseness-based 2ch BSS using the EM algorithm in reverberant environment, in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 147–150.

Kameoka, H., Goto, M., and Sagayama, S. (2006) Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelopes, in *IPSJ SIG Technical Reports*, vol. 2006-MUS-66-13, vol. 2006-MUS-66-13, pp. 77–84. In Japanese.

Kameoka, H. and Kashino, K. (2009) Composite autoregressive system for sparse source-filter representation of speech, in *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 2477–2480.

Kameoka, H., Sato, M., Ono, T., Ono, N., and Sagayama, S. (2012) Blind separation of infinitely many sparse sources, in *Proceedings of International Workshop on Acoustic Echo and Noise Control*.

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

Kitamura, D., Ono, N., Sawada, H., Kameoka, H., and Saruwatari, H. (2015) Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 276–280.

Kounades-Bastian, D., Girin, L., Alameda-Pineda, X., Gannot, S., and Horaud, R. (2016) A variational EM algorithm for the separation of time-varying convolutive audio mixtures. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (8), 1408 – 1423.

Lee, D.D. and Seung, H.S. (2000) Algorithms for non-negative matrix factorization, in *Proceedings of Neural Information Processing Systems*, vol. 13, vol. 13, pp. 556 – 562.

Leeuw, J.D. and Heiser, W.J. (1977) Convergence of correction matrix algorithms for multidimensional scaling, in *Geometric Representations of Relational Data*, Mathesis Press.

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

Nikunen, J. and Virtanen, T. (2014) Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22** (3), 727–739.

Nugraha, A.A., Liutkus, A., and Vincent, E. (2016) Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (9), 1652 – 1664.

Ono, N. (2011) Stable and fast update rules for independent vector analysis based on auxiliary function technique, in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 189–192.

Ozerov, A. and Févotte, C. (2010) Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, **18** (3), 550–563.

Ozerov, A., Févotte, C., Blouet, R., and Durrieu, J.L. (2011) Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 257–260.

Ozerov, A., Févotte, C., and Charbit, M. (2009) Factorial scaled hidden Markov model for polyphonic audio representation and source separation, in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 121–124.

Ozerov, A., Vincent, E., and Bimbot, F. (2012) A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*,

20 (4), 1118–1133.

Petersen, K.B. and Pedersen, M.S. (2005) The matrix cookbook. Version 3.

Pham, D.T., Servière, C., and Boumaraf, H. (2003) Blind separation of speech mixtures based on nonstationarity, in *Proceedings of International Conference on Information Sciences, Signal Processing and their Applications*, pp. II–73–II–76.

Sawada, H., Kameoka, H., Araki, S., and Ueda, N. (2013) Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (5), 971–982.

Thiemann, J. and Vincent, E. (2013) A fast EM algorithm for Gaussian model-based source separation, in *Proceedings of European Signal Processing Conference*.

Vincent, E., Arberet, S., and Gribonval, R. (2009) Underdetermined instantaneous audio source separation via local Gaussian modeling, in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 775 – 782.

Vincent, E. and Rodet, X. (2004) Underdetermined source separation with structured source priors, in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 327–332.

Yoshioka, T., Nakatani, T., Miyoshi, M., and Okuno, H.G. (2011) Blind separation and dereverberation of speech mixtures by joint optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, **19** (1), 69–84.