Audio source separation

Roland Badeau
roland.badeau@telecom-paris.fr

Contents

Roland Badeau roland.badeau@telecom-paris.fr

Voir Page 28

List of Figures

Acronyms

BSS *Blind source separation*

DTFT *Discrete time Fourier transform*

DUET *Degenerate unmixing estimation technique*

ICA *Independent component analysis*

IID *Independent and identically distributed*

JADE *Joint approximate diagonalization of eigenmatrices*

MDCT *Modified discrete cosine transform*

MMSE *Minimum mean square error*

MSE *Mean square error*

PDF *Probability density function*

PSD *Power spectral density*

SOBI *Second order blind identification*

STFT *Short time Fourier transform*

TF *Time-frequency*

WSS *Wide sense stationary*

Mathematical notation

$\mathbb{R}$ set of real numbers

$\mathbb{C}$ set of complex numbers

x (normal font, lower case) scalar

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

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

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

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

$.^\top$ transpose of a matrix

$.^H$ conjugate transpose of a matrix

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

trace(.) trace of a matrix

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

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

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

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

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

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

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

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

Chapter 1

Audio source separation

1 Introduction

Source separation is the art of estimating *source* signals, which are assumed statistically independent, from the observation of one or several *mixtures* of these signals. It is useful in many audio signal processing tasks, including *denoising* applications:

- separation of the instruments in polyphonic music;
- karaoke: remove the singer voice in music recordings;
- cocktail party problem: isolate the voice of the person you are speaking to from many other voices;
- suppression of vuvuzela in TV broadcasting of football matches during the 2010 FIFA world cup.

Besides, the separated audio tracks can be used for remixing purposes, possibly including transformations (e.g. pitch shifting, time scaling, etc.) or re-spatialization of the separated audio sources.

1.1 Typology of the mixture models

Formally, the observed data is made of M mixture signals $x_m(t)$, concatenated in a vector $\boldsymbol{x}(t)$. The unknowns are the K (possibly different from M) source signals $s_k(t)$, concatenated in a vector $\boldsymbol{s}(t)$.

The mixture is modeled as a function $\mathcal{A}$ which transforms the source signals $\boldsymbol{s}(t)$ into the mixture signals $\boldsymbol{x}(t)$. Generally, some simplifying assumptions are introduced regarding the mixture model [VVG18, chap. 1]:

- *Stationarity*: function $\mathcal{A}$ is translation invariant.
- *Linearity*: function $\mathcal{A}$ is a linear map.
- *Memory*:
  - Transformations that are both stationary and linear can be modeled with convolution products in the time domain (linear filtering).
  - The *memory* of such transformations corresponds to the length of the impulse response.
  - If there is no memory (i.e. the length is zero), the mixture is called *instantaneous* and $\mathcal{A}$ is characterized by a *mixing matrix* $\boldsymbol{A}$ (of dimension $M \times K$): $\boldsymbol{x}(t) = \boldsymbol{A}\,\boldsymbol{s}(t)$. Instantaneous mixture models are suitable e.g. for some biomedical applications (electroencephalography (EEG) or magnetoencephalography (MEG)), but generally not for audio applications, because of reverberation.

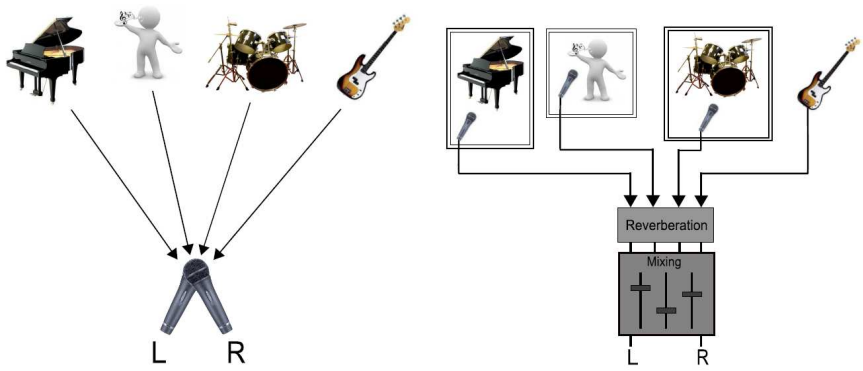
Depending on the respective values of M and K, the mixture may or may not be invertible:

- If $M = K$, the mixture is called *determined*: it is generally invertible.
- If $M > K$, the mixture is called *over-determined*: a unique solution can be found in the least squares sense.
- If $M < K$, the mixture is called *under-determined*: there are infinitely many solutions. Without additional information about the mixture or the source signals, it is impossible to retrieve the original sources from the mixture signals.

1.2 Instantaneous linear mixtures

Examples of instantaneous linear mixtures are given in Figure 1.1:

- In a real audio environment, an approximately instantaneous linear mixture can be obtained with the X-Y stereo recording technique, by putting two directional microphones at the same place, typically oriented at 90 degrees or more from each other (Figure 1.1-(a)). However the audio mixture obtained in this way is never perfectly instantaneous.
- Otherwise, truly instantaneous linear mixtures can of course be created artificially by using a mixing deck or a computer (Figure 1.1-(b)).



(a) Configuration stéréo XY. (b) Injection directe sur la table de mixage.

Figure 1.1: Instantaneous linear mixtures

1.3 Anechoic linear mixtures

Anechoic linear mixtures are a particular case of convolutive mixtures that can be recorded in an *anechoic chamber*: because the sound reflections on the room walls are greatly attenuated, every impulse response is formed only of a single pulse, characterized by its delay and its magnitude, which corresponds to the direct propagation path from every source to every microphone (Figure 1.2).

1.4 Convolutive mixtures

In the general case, audio mixtures are convolutive: in a room, the sound waves are reflected on the walls, so the impulse response is formed of infinitely many pulses, which correspond to the direct propagation path and the various reflections, whose density grows quadratically with time. This phenomenon is called *reverberation* (Figure 1.3-(a)). Convolutive mixtures can also be created artificially, e.g. to simulate a 3-D stereo sound sensation for the listener using headphones (*binaural* mixture, illustrated in Figure 1.3-(b)).

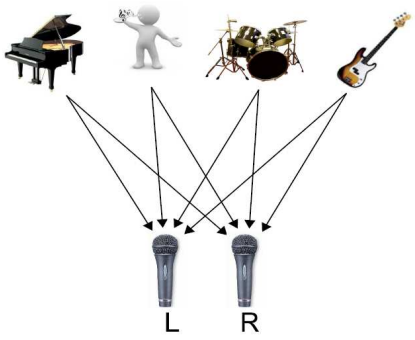
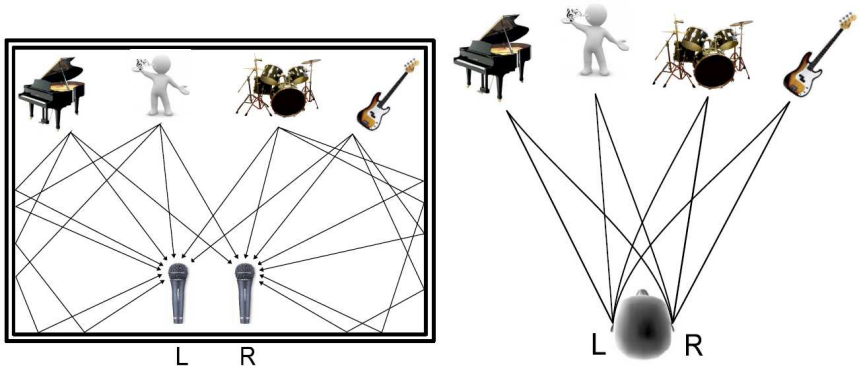


Figure 1.2: Anechoic linear mixtures



(a) Mélange convolutif.

(b) Mélange binaural.

Figure 1.3: Convolutive mixtures

2 Mathematical reminders

Because most source separation techniques involve probabilistic models, we first start with some mathematical reminders from probability theory and statistical signal processing.

2.1 Real random vectors

Let $\boldsymbol{x} \in \mathbb{R}^M$ denote a real random vector. In the rest of this document, we will use the following notation: $\phi[\boldsymbol{x}]$ (with square brackets) denotes a function of the distribution of the random vector $\boldsymbol{x}$, whereas a random variable defined as a function of $\boldsymbol{x}$ would be denoted $\psi(\boldsymbol{x})$ (with parentheses). In particular, we will consider:

- the mean vector: $\boldsymbol{\mu}_x = \mathbb{E}[\boldsymbol{x}] \in \mathbb{R}^M$ (where $\mathbb{E}$ denotes the *mathematical expectation*, a.k.a the *expected value*);
- the covariance matrix: $\boldsymbol{\Sigma}_{xx} = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top] \in \mathbb{R}^{M \times M}$, which is always symmetric (i.e. $\boldsymbol{\Sigma}_{xx}^\top = \boldsymbol{\Sigma}_{xx}$ where $.^\top$ denotes the transpose of a matrix) and positive semi-definite (i.e. $\forall \boldsymbol{v} \in \mathbb{R}^M, \boldsymbol{v}^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{v} \geq 0$);
- the characteristic function: $\phi_x(\boldsymbol{f}) = \mathbb{E}[e^{-2i\pi \boldsymbol{f}^\top \boldsymbol{x}}] \in L^\infty(\mathbb{R}^M)$ (where $L^\infty(\mathbb{R}^M)$ denotes the Lebesgue space of essentially bounded functions on $\mathbb{R}^M$);
- when the inverse Fourier transform of ϕ_x is a measurable function on $\mathbb{R}^M$, $p(\boldsymbol{x}) = \int_{\mathbb{R}} \phi_x(\boldsymbol{f}) e^{+2i\pi \boldsymbol{f}^\top \boldsymbol{x}} d\boldsymbol{f}$ is called the *Probability density function* (PDF) of the random vector $\boldsymbol{x}$.

Some of the oldest source separation methods are based on the notion of *cumulants*. The cumulants of the random vector $\boldsymbol{x}$ will be denoted $\kappa^n_{k_1\ldots k_n}[\boldsymbol{x}] \in \mathbb{R}$ for all orders $n \in \mathbb{N}$ and entries $k_i \in \{1 \ldots M\}$, and they are defined as the coefficients of the Taylor expansion of the *cumulant generating function*, which is the natural logarithm of the characteristic function: when ϕ_x is an analytic function, we can write

$$\ln(\phi_x(\boldsymbol{f})) = \sum_{n=1}^{+\infty} \frac{(-2\imath\pi)^n}{n!} \sum_{k_1=1}^{M} \sum_{k_n=1}^{M} \kappa^n_{k_1\ldots k_n}[\boldsymbol{x}] f_{k_1} \ldots f_{k_n}.$$

The cumulants satisfy the following properties:

- $\forall n \in \mathbb{N}^*$, $\kappa^n[\boldsymbol{x}]$ is an n-th order tensor of coefficients $\kappa^n_{k_1\ldots k_n}[\boldsymbol{x}]$;
- $\kappa^1[\boldsymbol{x}]$ is the mean vector $\boldsymbol{\mu}_x$ and $\kappa^2[\boldsymbol{x}]$ is the covariance matrix $\boldsymbol{\Sigma}_{xx}$;
- If the PDF $p(\boldsymbol{x})$ is symmetric ($p(-\boldsymbol{x}) = p(\boldsymbol{x})$), then $\kappa^n[\boldsymbol{x}] = 0$ for any odd value n;
- The ratio between the fourth order cumulant $\kappa^4_{k,k,k,k}[\boldsymbol{x}]$ and the squared variance $(\kappa^2_{k,k}[\boldsymbol{x}])^2$ plays a special role in independent component analysis (*cf.* Section 3.2). It is called the *excess kurtosis*.

2.2 Real Gaussian random vectors

Among all probability distributions with well-defined cumulants of all orders, the Gaussian distribution is the one such that all cumulants of order $n > 2$ are zero. The characteristic function of a Gaussian random vector $\boldsymbol{x} \in \mathbb{R}^M$ can thus be expressed as

$$\phi_x(\boldsymbol{f}) = \exp\left(-2\imath\pi \boldsymbol{f}^\top \boldsymbol{\mu}_x - 2\pi^2 \boldsymbol{f}^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{f}\right).$$

When the covariance matrix $\boldsymbol{\Sigma}_{xx}$ is invertible, then the PDF is defined as:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{M}{2}} \det(\boldsymbol{\Sigma}_{xx})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top \boldsymbol{\Sigma}_{xx}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x)\right).$$

2.3 WSS vector processes

A discrete vector process is a sequence of random vectors $\boldsymbol{x}(t) \in \mathbb{R}^M$ indexed by time $t \in \mathbb{Z}$. A second order vector process is a discrete vector process with well-defined second order moments. Finally, a *Wide sense stationary* (WSS) vector process $\boldsymbol{x}(t)$ is a second order vector process whose cumulants of orders 1 and 2 are invariant under any translation of time:

- $\mathbb{E}[\boldsymbol{x}(t)] = \boldsymbol{\mu}_x \ \forall t \in \mathbb{Z}$ where $\boldsymbol{\mu}_x \in \mathbb{R}^M$ is the *mean vector* of the vector process $\boldsymbol{x}(t)$;
- $\forall t \in \mathbb{Z}$, $\mathbb{E}\left[(\boldsymbol{x}(t+\tau) - \boldsymbol{\mu}_x)(\boldsymbol{x}(t) - \boldsymbol{\mu}_x)^\top\right] = \boldsymbol{R}_{xx}(\tau)$, where $\forall \tau \in \mathbb{Z}$ $\boldsymbol{R}_{xx}(\tau) \in \mathbb{R}^{M \times M}$ defines the *autocovariance function* of the vector process $\boldsymbol{x}(t)$. When $\tau = 0$, $\boldsymbol{R}_{xx}(0) = \boldsymbol{\Sigma}_{xx}$ is the covariance matrix of the random vector $\boldsymbol{x}(t)$ $\forall t \in \mathbb{Z}$, and as such it is symmetric and positive semi-definite.

Finally, given two jointly WSS vector processes $\boldsymbol{x}(t) \in \mathbb{R}^M$ and $\boldsymbol{y}(t) \in \mathbb{R}^N$ of mean zero, we define their *intercovariance function* $\boldsymbol{R}_{xy}(\tau) \in \mathbb{R}^{M \times N}$:

$$\forall \tau \in \mathbb{Z},\ \boldsymbol{R}_{xy}(\tau) = \mathbb{E}\left[\boldsymbol{x}(t+\tau)\boldsymbol{y}(t)^\top\right].$$

When the *Discrete time Fourier transform* (DTFT) of the autocovariance function $\boldsymbol{R}_{xx}(\tau)$ of a WSS vector process $\boldsymbol{x}(t)$ is a measurable function $\boldsymbol{S}_{xx}(\nu) \in \mathbb{C}^{M \times M}$, this function is called the *Power spectral density* (PSD) of $\boldsymbol{x}(t)$:

$$\forall \nu \in \mathbb{R},\ \boldsymbol{S}_{xx}(\nu) = \sum_{\tau \in \mathbb{Z}} \boldsymbol{R}_{xx}(\tau) e^{-2\imath\pi\nu\tau}.$$

The PSD is always periodic of period 1, and $\forall \nu \in \mathbb{R}$, matrix $\boldsymbol{S}_{xx}(\nu)$ is always Hermitian symmetric (i.e. $\boldsymbol{S}_{xx}(\nu)^H = \boldsymbol{S}_{xx}(\nu)$ where $.^H$ denotes the conjugate transpose of a matrix) and positive semi-definite (i.e. $\forall \boldsymbol{v} \in \mathbb{C}^M$, $\boldsymbol{v}^H \boldsymbol{S}_{xx}(\nu) \boldsymbol{v} \geq 0$).

2.4 Information theory

Information theory is a fundamental tool in *blind source separation* (*cf.* Section 3.1), because it makes it possible to measure the amount of information shared between several random variables.

We first consider the notion of *entropy*, which measures the degree of uncertainty in a probability distribution. For a discrete random variable x with probability distribution p, the *Shannon entropy* is defined as $\mathbb{H}[x] = -\mathbb{E}[\ln(p(x))]$. It is always a non-negative real number. The higher this number, the more "uncertain" the outcome of x is. For a continuous random vector $\boldsymbol{x} \in \mathbb{R}^M$ with PDF $p(\boldsymbol{x})$, the *differential entropy* is defined in the same way: $\mathbb{H}[\boldsymbol{x}] = -\mathbb{E}[\ln(p(\boldsymbol{x}))]$. However the differential entropy $\mathbb{H}[\boldsymbol{x}]$ is not necessarily non-negative.

For continuous random vectors $\boldsymbol{x} \in \mathbb{R}^M$, the *Kullback-Leibler divergence* measures the degree of dissimilarity between two probability distributions characterized by their PDFs p and q:

$$D_{KL}(p\|q) = \int_{\boldsymbol{x}\in\mathbb{R}^M} p(\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x}.$$

As a *divergence*, it is always nonnegative, and $D_{KL}(p\|q) = 0$ if and only if $p = q$. However, the Kullback-Leibler divergence is not a *distance*, because it is not symmetric (in general $D_{KL}(p\|q) \neq D_{KL}(q\|p)$), and it does not satisfy the triangle inequality.

Finally, the *mutual information* measures the mutual dependence between several random variables. For instance if $\boldsymbol{x} \in \mathbb{R}^M$ is a continuous random vector, then the mutual information between the entries of $\boldsymbol{x}$ is defined as:

$$\mathbb{I}[\boldsymbol{x}] = \mathbb{E}\left[\ln\left(\frac{p(\boldsymbol{x})}{p(x_1)\dots p(x_M)}\right)\right] = D_{KL}(p(\boldsymbol{x})\|p(x_1)\dots p(x_M)).$$

Since D_{KL} is a divergence, $\mathbb{I}[\boldsymbol{x}]$ is always nonnegative, and $\mathbb{I}[\boldsymbol{x}] = 0$ if and only if $p(\boldsymbol{x}) = p(x_1)\dots p(x_M)$, i.e. if and only if the random variables $x_1 \dots x_M$ are mutually independent. The mutual information is related to the differential entropy through the equality

$$\mathbb{I}[\boldsymbol{x}] = \left(\sum_{m=1}^{M} \mathbb{H}[x_m]\right) - \mathbb{H}[\boldsymbol{x}]. \tag{1.1}$$

In *independent component analysis*, the mutual information is an objective function to be minimized, in order to make several random variables as independent as possible (*cf.* Section 3.2.2). Equation (1.1) shows that minimizing $\mathbb{I}[\boldsymbol{x}]$ is equivalent to minimizing the sum of the individual entropies $\mathbb{H}[x_m]$ when the joint entropy $\mathbb{H}[\boldsymbol{x}]$ is fixed.

3 Linear instantaneous mixtures

Even though we have seen in Section 1.2 page 6 that the linear instantaneous mixture model cannot accurately represent real acoustic mixtures, the oldest separation techniques which paved the way for modern audio source separation methods are based on this model. We thus first address this over-simplified mixture model, which will permit us to introduce several useful concepts and methods, that will then be extended to the more realistic convolutive mixture model in Section 4.

3.1 Blind source separation (BSS) model

Blind source separation (BSS) techniques [Car98] assume that we know very little about the source signals: they are only assumed to be statistically independent. This is the case for instance in many denoising applications, where the signal of interest (e.g. speech) is independent from the source of noise (e.g. background environmental noise). This hypothesis is the funding principle of most multichannel source separation methods.

Actually, BSS methods also rely on a generative source model, but this model is chosen as little informative as possible: the samples of each source signal are assumed *Independent and identically distributed* (IID). The IID source model thus ignores any temporal dynamics (i.e. power variations over time), or spectral dynamics (i.e. temporal correlations), that might be present in the source signals:

Definition 1 (IID source model) *We consider K independent source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$. For all $k \in \{1 \dots K\}$, s_k is modeled as an IID random process: the samples $s_k(t)$ are independent random variables, of same probability distribution p_k (which depends on source k).*

The possibility of performing source separation in such a blind way may seem to be an incredible feat. Actually, the trick is that, contrary to the source model, the mixture model is *very* constraining: at first we will only consider linear instantaneous mixtures, characterized by a mixing matrix $\boldsymbol{A}$:

Definition 2 (Linear instantaneous mixture model) *We consider K source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$. Then the samples of the M mixture signals $x_m(t) \in \mathbb{R}$ for $m \in \{1 \dots M\}$ are defined as the entries of the M-dimensional vector*

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t), \tag{1.2}$$

where $\boldsymbol{A} \in \mathbb{R}^{M \times K}$ is called the mixing matrix*, and $\boldsymbol{s}(t)$ is the K-dimensional vector of entries $s_k(t)$ for $k \in \{1 \dots K\}$.*

Given the source model in Definition 1 and the mixture model in Definition 2, the purpose of BSS is to estimate the source signals $s_k(t)$ given the observed mixture signals $x_m(t)$, *without knowing* the mixing matrix $\boldsymbol{A}$. When the mixture is *determined* ($M = K$) and matrix $\boldsymbol{A}$ is invertible, we will show that this is generally feasible.

3.1.1 Identifiability

Suppose that the mixture is determined ($M = K$) . Before investigating how source separation can be performed, we first need to study the *identifiability* of the linear instantaneous BSS model: is it really possible to retrieve both the source signals $s_k(t)$ and the mixing matrix $\boldsymbol{A}$ from only the observed mixture signals $x_m(t)$?

Clearly, if $\boldsymbol{P}$ is a permutation matrix (i.e. it has a unique 1 entry in each row and each column, all other entries being 0), then matrix $\tilde{\boldsymbol{A}} = \boldsymbol{A}\,\boldsymbol{P}^{-1}$ and vector $\tilde{\boldsymbol{s}}(t) = \boldsymbol{P}\boldsymbol{s}(t)$ lead to the same observations $\boldsymbol{x}(t)$, while satisfying all the properties of the linear instantaneous BSS model in Definitions 1 and 2. So the source signals can only be retrieved up to a permutation: at best we can retrieve the source signals, but we cannot *identify* them.

In the same way, if $\boldsymbol{D}$ is an invertible diagonal matrix, then matrix $\tilde{\boldsymbol{A}} = \boldsymbol{A}\,\boldsymbol{D}^{-1}$ and vector $\tilde{\boldsymbol{s}}(t) = \boldsymbol{D}\boldsymbol{s}(t)$ lead to the same observations $\boldsymbol{x}(t)$, while satisfying all the properties of the linear instantaneous BSS model in Definitions 1 and 2. Therefore the source signals can only be retrieved up to a multiplicative factor (which in most applications is not a problem: e.g. audio signals are generally scaled during playback).

So the linear instantaneous BSS model in Definitions 1 and 2 has at least *permutation* and *scale* indeterminacies. Actually, it can be proved that there is no other one. These two indeterminacies are summarized by the concept of *non-mixing matrices*:

Definition 3 (Non-mixing matrix) *A matrix $\boldsymbol{C} \in \mathbb{R}^{K \times K}$ is* non-mixing *if and only if it has a unique non-zero entry in each row and each column.*

A non-mixing matrix can always be decomposed as the product of a permutation matrix and an invertible diagonal matrix.

3.1.2 Linear separation of sources

When the mixture is linear instantaneous, it may seem natural to estimate the source signals as linear instantaneous combinations of the mixture signals:

$$\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t), \tag{1.3}$$

where the entries of vector $\boldsymbol{y}(t)$ are the source signal estimates, and $\boldsymbol{B} \in \mathbb{R}^{K \times M}$ is referred to as the *separation matrix*. Then the source separation problem amounts to finding an optimal separation matrix.

Linear source separation is generally feasible in the case of determined and over-determined mixtures:

- if $M = K$ and if matrix $\boldsymbol{A}$ is invertible, then the separation matrix $\boldsymbol{B} = \boldsymbol{A}^{-1}$ leads to $\boldsymbol{y}(t) = \boldsymbol{s}(t)$;
- more generally, if $M \geq K$ and if matrix $\boldsymbol{A}$ has full rank, then the separation matrix $\boldsymbol{B} = \boldsymbol{A}^{\dagger}$ leads to $\boldsymbol{y}(t) = \boldsymbol{s}(t)$, where $.^{\dagger}$ denotes the matrix the pseudo-inverse: $\boldsymbol{A}^{\dagger} = (\boldsymbol{A}^{\top}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}$, which is such that $\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{I}_K$.

However, in the under-determined case ($M < K$), linear source separation is generally not feasible (*cf.* Section 5).

3.2 Independent component analysis (ICA)

Independent component analysis (ICA) [CJ10] is a linear source separation technique which consists in looking for a separation matrix $\boldsymbol{B}$ that makes the signals $y_k(t)$ independent.

Since $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t)$ (equation (1.2)) and $\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t)$ (equation (1.3)), we have $\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{s}(t)$ with $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}$. According to the identifiability analysis in Section 3.1.1, the BSS problem is solved if and only if matrix $\boldsymbol{C}$ is non-mixing (*cf.* Figure 1.4).

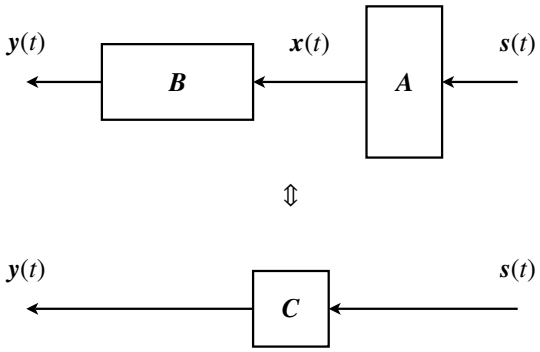


Figure 1.4: Identifiability theorem: signals $y_k(t)$ are independent if and only if matrix $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}$ is non-mixing

The following identifiability theorem due to P. Comon [Com94] proves the feasibility of ICA under mild conditions about the source signals:

Theorem 1 (Identifiability theorem) *Consider the linear instantaneous BSS model in Definitions 1 and 2 in the determined case ($M = K$). Among the K IID sources s_k, suppose that at most one is Gaussian-distributed. Let $\boldsymbol{C} \in \mathbb{R}^{K\times K}$ and $\forall t \in \mathbb{Z}$, $\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{s}(t)$. Then the random processes $y_k(t)$ for $k \in \{1 \dots K\}$ are independent if and only if matrix $\boldsymbol{C}$ is non-mixing.*

Theorem 1 proves that finding a separation matrix $\boldsymbol{B}$ that makes signals $y_k(t)$ independent solves the BSS problem: the estimated signals $y_k(t)$ are equal to the source signals $s_k(t)$ up to permutation and scale indeterminacies.

Here, pay attention to the non-Gaussianity assumption in Theorem 1: in Section 3.2.1, we will show that indeed, if two (or more) sources are Gaussian, then the BSS problem cannot be solved.

3.2.1 Whitening

Independent component analysis can be performed in two steps; the first one consists in *whitening*[1], i.e. *decorrelating* the observed mixture signals. Remember that independence implies decorrelation, but decorrelation does generally not imply independence. Therefore the second step will consist in making the whitened signals independent.

To simplify the problem, we will address the case of determined mixtures $M = K$ (even though whitening could also be performed in the over-determined case $M > K$), and we will assume that matrix $\boldsymbol{A}$ is invertible and that the source signals are centered: $\mathbb{E}[\boldsymbol{s}(t)] = \boldsymbol{0}$ (which is always the case of audio signals).

To further simplify, we will focus on the *canonical BSS problem*: without loss of generality, we will assume that the random vectors $\boldsymbol{s}(t)$ are spatially white, i.e. their covariance matrix is $\boldsymbol{\Sigma}_{ss} = \mathbb{E}[\boldsymbol{s}(t)\boldsymbol{s}(t)^\top] = \boldsymbol{I}_K$. Indeed, since the source signals are independent, we already know that matrix $\boldsymbol{\Sigma}_{ss}$ is diagonal; since in addition the source signals can only be retrieved up to a multiplicative factor, we can also assume without loss of generality that the diagonal entries of matrix $\boldsymbol{\Sigma}_{ss}$ are 1.

Since $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t)$ (equation (1.2)), the covariance matrix of the mixture vectors $\boldsymbol{x}(t)$ is $\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}\boldsymbol{\Sigma}_{ss}\boldsymbol{A}^\top = \boldsymbol{A}\,\boldsymbol{A}^\top$: we say that $\boldsymbol{A}$ is a *matrix square root* of $\boldsymbol{\Sigma}_{xx}$. This property is interesting because $\boldsymbol{\Sigma}_{xx}$ can be estimated from the

[1] Whitening is performed in the spatial domain, i.e. over channels, not in the time domain, i.e. over time samples.

observed data, and it carries information about the mixing matrix $\boldsymbol{A}$. Unfortunately, we will see that this property is not sufficient to fully characterize $\boldsymbol{A}$. Nevertheless, it allows us to make a first step towards the estimation of $\boldsymbol{A}$. For the moment, just note that since matrix $\boldsymbol{A}$ is invertible, matrix $\boldsymbol{\Sigma}_{xx}$ is also invertible, thus positive definite.

The whitening of the mixture signals can then be performed as follows [CS93]:

- Since matrix $\boldsymbol{\Sigma}_{xx}$ is positive definite, the spectral theorem in matrix theory shows us that it is diagonalizable in an orthonormal basis: there is an orthonormal matrix $\boldsymbol{Q} \in \mathbb{R}^{K\times K}$ (i.e. such that $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\top$), and a diagonal matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{K\times K}$ with positive diagonal entries, such that
$$\boldsymbol{\Sigma}_{xx} = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top. \tag{1.4}$$
- Then let $\boldsymbol{S} = \boldsymbol{Q}\,\boldsymbol{\Lambda} \in \mathbb{R}^{K\times K}$; matrix $\boldsymbol{S}$ is also a matrix square root of $\boldsymbol{\Sigma}_{xx}$, since $\boldsymbol{S}\,\boldsymbol{S}^\top = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top = \boldsymbol{\Sigma}_{xx}$.
- Finally, let
$$\boldsymbol{W} = \boldsymbol{S}^{-1} \tag{1.5}$$
and
$$\forall t \in \mathbb{Z},\ \boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t). \tag{1.6}$$

Then the random vector process $\boldsymbol{z}(t)$ is spatially white, in the sense that on the one hand it is centered: $\mathbb{E}[\boldsymbol{z}(t)] = \boldsymbol{0}$ (since $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{A}\boldsymbol{s}(t)$ and $\mathbb{E}[\boldsymbol{s}(t)] = \boldsymbol{0}$), and on the other hand its covariance matrix is $\boldsymbol{\Sigma}_{zz} = \boldsymbol{W}\boldsymbol{\Sigma}_{xx}\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{W}^\top = \boldsymbol{I}_K$). Matrix $\boldsymbol{W}$ will thus be referred to as the *whitening matrix* and $\boldsymbol{z}(t)$ is the *whitened data*.

Then let us define matrix $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{A}$. We have $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{W}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{\Sigma}_{xx}\boldsymbol{W}^\top = \boldsymbol{I}_K$, therefore $\boldsymbol{U}$ is an orthonormal matrix. In particular, $|\det(\boldsymbol{U})| = 1$: if $\det(\boldsymbol{U}) = 1$, $\boldsymbol{U}$ is a *rotation matrix*, otherwise if $\det(\boldsymbol{U}) = -1$, $\boldsymbol{U}$ is a *reflection matrix*. However, remember that the source signals can only be retrieved up to a multiplicative factor, which might as well be negative. Therefore, by changing the sign of the k-th source signal $s_k(t)$, the product $\boldsymbol{A}\boldsymbol{s}(t)$ is left unchanged by changing the sign of the k-th column of matrix $\boldsymbol{A}$, which changes the sign of $\det(\boldsymbol{U})$. Therefore, without loss of generality, we can assume that $\boldsymbol{U}$ *is a rotation matrix* ($\det(\boldsymbol{U}) = 1$).

Finally, let
$$\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{z}(t). \tag{1.7}$$

Then $\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{W}\boldsymbol{x}(t) = (\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{W}(\boldsymbol{A}\boldsymbol{s}(t)) = \boldsymbol{s}(t)$. Therefore matrix $\boldsymbol{B} = \boldsymbol{U}^\top\boldsymbol{W}$ is a separation matrix. In practice of course, matrix $\boldsymbol{A}$ is unknown, thus so is matrix $\boldsymbol{U}$. But it remains that ICA can be performed in two steps (*cf.* Figure 1.5):

Step 1 : compute the whitening matrix $\boldsymbol{W}$ and the whitened data $\boldsymbol{z}(t)$ from an estimate of the covariance matrix $\boldsymbol{\Sigma}_{xx}$;

Step 2 : look for a rotation matrix $\boldsymbol{U}$ such that the entries of vector $\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{z}(t)$ are independent.

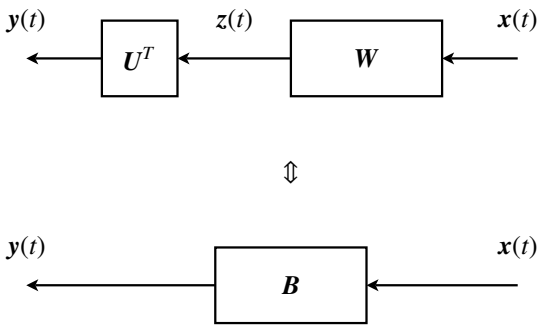


Figure 1.5: Pre-whitening for independent component analysis: $\boldsymbol{B} = \boldsymbol{U}^\top\boldsymbol{W}$ where $\boldsymbol{U}$ is a rotation matrix

To summarize, the whiteness property (based on second order cumulants) determines matrix $\boldsymbol{W}$ and leaves the rotation matrix $\boldsymbol{U}$ unknown.

Note that in the Gaussian case, decorrelation implies independence. Therefore if the source signals are Gaussian-distributed, then the whitened signals $z_k(t)$ are independent, and $\boldsymbol{U}$ cannot be determined. This explains the assumption made in Theorem 1 page 11 that at most one source can be Gaussian-distributed (if they are two of them, they cannot be separated).

Therefore if we want to determine rotation $\boldsymbol{U}$, we will need to explicitly exploit the non-Gaussianity of the source signals. To do so, we will characterize the independence property by using cumulants of order greater than 2 [CS93].

3.2.2 Contrast functions

In Section 2.4, we have introduced the concept of *mutual information*: the mutual information between several random variables is always non-negative, and it is zero if and only if these random variables are independent. Therefore the mutual information between the entries of vector $\boldsymbol{y}(t)$ can be used as an objective function to be minimized in order to perform ICA.

More generally, the concept of *contrast functions* has been introduced in order to formulate ICA as an optimization problem, the mutual information being only one example of such functions. Formally, still in the determined case ($M = K$), Theorem 1 leads to the following definition of *contrast functions* [Car98]:

Definition 4 (Contrast function) *For all $k \in \{1 \dots K\}$, we consider source signals $s_k(t)$ as defined in Definition 1. Then a function $\phi : \mathbb{R}^K \to \mathbb{R}$ is a* contrast function *when $\phi[\boldsymbol{C}\boldsymbol{s}(t)] \geq \phi[\boldsymbol{s}(t)]$ for any matrix $\boldsymbol{C} \in \mathbb{R}^{K \times K}$, and $\phi[\boldsymbol{C}\boldsymbol{s}(t)] = \phi[\boldsymbol{s}(t)]$ if and only if matrix $\boldsymbol{C}$ is non-mixing.*

Following Definition 4 and considering linear instantaneous mixtures $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t)$ (equation (1.2)) as in Definition 2 page 10, linear source separation $\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t)$ (equation (1.3) page 10), and matrix $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}$, the BSS problem is solved by minimizing the contrast function $\phi[\boldsymbol{y}(t)]$ with respect to (w.r.t.) the separation matrix $\boldsymbol{B}$, or w.r.t. the rotation matrix $\boldsymbol{U}$ if the data has been whitened. Among all possible contrast functions, the mutual information $\phi_{IM}[\boldsymbol{y}(t)] = \mathbb{I}[\boldsymbol{y}(t)]$ is considered as the *canonical* contrast function.

If the observed data has already been whitened, it is possible to consider only *orthogonal contrast functions*, which are such that ICA is performed by minimizing an orthogonal contrast function subject to the constraint $\mathbb{E}[\boldsymbol{y}(t)\boldsymbol{y}(t)^\top] = \boldsymbol{I}_K$. For instance, by using equation (1.1) page 9, it can be shown that an orthogonal contrast function associated to the mutual information is $\phi^\circ_{IM}[\boldsymbol{y}(t)] = \sum_{k=1}^K H(y_k(t))$, because $H(\boldsymbol{y}(t))$ is left unchanged under the constraint $\mathbb{E}[\boldsymbol{y}(t)\boldsymbol{y}(t)^\top] = \boldsymbol{I}_K$.

In practice, the orthogonal contrast function ϕ°_{IM} can be expressed as a function of the cumulants of $\boldsymbol{y}(t)$, and approximated by considering only the cumulants up to order 4 [CS93]:

$$\phi^\circ_{ICA}[\boldsymbol{y}(t)] = \sum_{ijkl \neq iiii} (\kappa^4_{ijkl}[\boldsymbol{y}(t)])^2. \tag{1.8}$$

Then the minimization of ϕ°_{ICA} with respect to the rotation matrix $\boldsymbol{U}$ can e.g. be performed by factorizing $\boldsymbol{U}$ as a product of Givens rotations (i.e. 2-dimensional rotation matrices parameterized by a single angle in $[0, 2\pi]$, which are applied iteratively to every pair of entries of vector $\boldsymbol{y}(t)$), and by performing a coordinate descent, also known as (a.k.a.) Jacobi technique, w.r.t. the angles of these Givens rotations [CJ10].

Compared to equation (1.8), the independence can also be tested on a smaller subset of cumulants, as

$$\phi^\circ_{JADE}[\boldsymbol{y}(t)] = \sum_{ijkl \neq ijkk} (\kappa^4_{ijkl}[\boldsymbol{y}(t)])^2. \tag{1.9}$$

The motivation for using this specific subset is that $\phi^\circ_{JADE}[\boldsymbol{y}(t)]$ can also be seen as a joint diagonalization criterion[2]. This approach leads to the celebrated *Joint approximate diagonalization of eigenmatrices* (JADE) method [CS93] summarized in Algorithm 1.

[2] Joint diagonalization of matrices will be addressed in Section 3.3.1.

Algorithm 1 JADE method

Estimation of the covariance matrix $\boldsymbol{\Sigma}_{xx}$ and diagonalization: $\boldsymbol{\Sigma}_{xx} = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top$ (equation (1.4))
Computation of $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}$ and of the whitening matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$ (equation (1.5))
Data whitening: $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ (equation (1.6))
Estimation of $\boldsymbol{U}$ by minimizing the contrast function ϕ°_{JADE} in equation (1.9)
Estimation of source signals via $\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{z}(t)$ (equation (1.7))

3.3 Second order methods

The JADE method is dedicated to the linear instantaneous BSS model in Definitions 1 and 2 page 10. It makes use of higher order statistics (i.e. of order greater than 2), because the identifiability of the model requires that at most one source signal be Gaussian distributed (*cf.* Theorem 1 page 11). However, it is well known that the estimating higher order statistics is more sensitive (e.g. in terms of mean square error) than estimating second order statistics. Therefore it would be interesting to develop source separation methods that only make use of second order statistics. That will require to relax the source model in Definition 1, so that the model become identifiable from its second order statistics only. In Sections 3.3.1 and 3.3.2, we will show two different ways of relaxing the source model.

3.3.1 Temporal coherence of source signals

In this section, we keep the same mixture model as in Definition 2, and we consider the source model in Definition 1 in the determined case ($M = K$), except that each source signal $s_k(t)$ is no longer assumed to be IID, but rather WSS, with a non-flat power spectral density. Therefore, as in Definition 1, the samples $s_k(t)$ for $t \in \mathbb{Z}$ can still follow the same distribution p_k, but now they are assumed to be mutually dependent (the source model is relaxed by removing the first "I" of "IID"):

Definition 5 (WSS source model) *We consider K independent source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$, concatenated in a vector $\boldsymbol{s}(t)$. For all $k \in \{1 \dots K\}$, s_k is modeled as a centered WSS random process. So $\boldsymbol{s}(t)$ is a WSS vector process of mean $\mathbb{E}[\boldsymbol{s}(t)] = \boldsymbol{0}$ and of autocovariance function $\boldsymbol{R}_{ss}(\tau) = \mathbb{E}\left[\boldsymbol{s}(t+\tau)\boldsymbol{s}(t)^\top\right]$.*

Since the source signals are independent, $\forall \tau \in \mathbb{Z}$ the covariance matrix $\boldsymbol{R}_{ss}(\tau)$ is diagonal: $\boldsymbol{R}_{ss}(\tau) = \mathrm{diag}(r_{s_k}(\tau))$, where $r_{s_k}(\tau) \in \mathbb{R}$ is the autocovariance function of the scalar WSS process $s_k(t)$, and diag(.) denotes a diagonal matrix formed from a vector of diagonal coefficients.

As in Section 3.2.1 page 11, we can still consider the canonical BSS problem and assume that $\boldsymbol{\Sigma}_{ss} = \boldsymbol{R}_{ss}(0) = \boldsymbol{I}_K$. Then, still as in Section 3.2.1, we can *spatially* whiten the mixture signals:

- compute a matrix square root $\boldsymbol{S}$ of $\boldsymbol{\Sigma}_{xx}$;
- compute $\boldsymbol{W} = \boldsymbol{S}^{-1}$ and the whitened data $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ $\forall t \in \mathbb{Z}$.

Again, since $\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}\,\boldsymbol{A}^\top$, matrix $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{A}$ is a rotation matrix. The novelty, compared with the mathematical developments in Section 3.2.1, is that we can now consider matrices $\boldsymbol{R}_{zz}(\tau)$ $\forall \tau \in \mathbb{Z}$, since they are no longer assumed to be zero. On the contrary, we have $\forall \tau \in \mathbb{Z}$, $\boldsymbol{R}_{zz}(\tau) = \boldsymbol{W}\boldsymbol{R}_{xx}(\tau)\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{A}\boldsymbol{R}_{ss}(\tau)\boldsymbol{A}^\top\boldsymbol{W}^\top = \boldsymbol{U}\boldsymbol{R}_{ss}(\tau)\boldsymbol{U}^\top$. This equation shows that matrices $\boldsymbol{R}_{zz}(\tau)$ are jointly diagonalized by the same set of eigenvectors, which are the columns of matrix $\boldsymbol{U}$, the eigenvalues being the diagonal entries of matrices $\boldsymbol{R}_{ss}(\tau)$.

Therefore the estimation of matrix $\boldsymbol{U}$ will no longer require the use of higher order statistics: $\boldsymbol{U}$ can be uniquely determined as the only rotation matrix (up to a non-mixing matrix) that jointly diagonalizes matrices $\boldsymbol{R}_{zz}(\tau)$ for different values of τ, as shown by the following theorem:

Theorem 2 (Unicity theorem) *Let us consider a set of matrices $\boldsymbol{R}_{zz}(\tau) \in \mathbb{R}^{K\times K}$ indexed by $\tau \in \mathbb{Z}$, of the form $\boldsymbol{R}_{zz}(\tau) = \boldsymbol{U}\boldsymbol{R}_{ss}(\tau)\boldsymbol{U}^\top$, where matrix $\boldsymbol{U} \in \mathbb{R}^{K\times K}$ is orthonormal and matrices $\boldsymbol{R}_{ss}(\tau) \in \mathbb{R}^{K\times K}$ are diagonal: $\boldsymbol{R}_{ss}(\tau) = \mathrm{diag}(r_{s_k}(\tau))$. Then $\boldsymbol{U}$ is unique (up to a non-mixing matrix) if and only if $\forall k \neq l \in \{1 \dots K\}$, there is $\tau \in \mathbb{Z}$ such that $r_{s_k}(\tau) \neq r_{s_l}(\tau)$.*

In order to compute matrix $\boldsymbol{U}$, we can thus use any joint diagonalization method [CS96]. For instance, we can numerically minimize the following objective function:

$$J(\boldsymbol{U}) = \sum_{\tau} \|\boldsymbol{U}^\top \boldsymbol{R}_{zz}(\tau)\boldsymbol{U} - \mathrm{diag}(\boldsymbol{U}^\top \boldsymbol{R}_{zz}(\tau)\boldsymbol{U})\|_F^2 \tag{1.10}$$

where $\|.\|_F$ denotes the Frobenius norm of a matrix (i.e. the Euclidean norm of a vector made of all its entries) and diag(.) denotes a diagonal matrix formed from a matrix with same diagonal entries. The criterion $J(\boldsymbol{U})$ is zero if and only if all matrices $\boldsymbol{U}^\top \boldsymbol{R}_{zz}(\tau)\boldsymbol{U}$ are diagonal.

As in Section 3.2.2 page 13, this minimization can be performed by factorizing $\boldsymbol{U}$ as a product of Givens rotations, and by performing a coordinate descent w.r.t. the angles of these Givens rotations [CJ10]. The resulting BSS method is known as the *Second order blind identification* (SOBI) technique [BAMCM97], and summarized in Algorithm 2.

Algorithm 2 SOBI method for WSS sources

Estimation of the covariance matrix $\boldsymbol{\Sigma}_{xx}$ and diagonalization: $\boldsymbol{\Sigma}_{xx} = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top$ (equation (1.4))
Computation of $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}$ and of the whitening matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$ (equation (1.5))
Data whitening: $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ (equation (1.6))
Estimation of covariance matrices $\boldsymbol{R}_{zz}(\tau)$ for various delays τ
Approximate joint diagonalization of matrices $\boldsymbol{R}_{zz}(\tau)$ in a common basis $\boldsymbol{U}$ by minimizing (1.10)
Estimation of source signals via $\boldsymbol{y}(t) = \boldsymbol{U}^\top \boldsymbol{z}(t)$ (equation (1.7))

3.3.2 Non-stationarity of source signals

In this section, we keep the same mixture model as in Definition 2, and we consider the source model in Definition 1 in the determined case ($M = K$), except that each source signal $s_k(t)$ is no longer assumed to be IID, but rather non-stationary. More precisely, as in Definition 1, the samples $s_k(t)$ for $t \in \mathbb{Z}$ can still be assumed independent, but now they no longer follow the same distribution p_k (the source model is relaxed by removing the last letters "ID" of "IID"):

Definition 6 (Non-stationary source model) *We consider K independent source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$, concatenated in a vector $\boldsymbol{s}(t)$. For all $k \in \{1 \dots K\}$, s_k is modeled as a centered random process with uncorrelated samples $s_k(t)$ for $t \in \mathbb{Z}$, of time-varying variance $\sigma_k^2(t)$. So $\boldsymbol{s}(t)$ is a random vector process of mean $\mathbb{E}[\boldsymbol{s}(t)] = \boldsymbol{0}$ and of time-varying covariance matrix $\boldsymbol{\Sigma}_{ss}(t) = \mathbb{E}\left[\boldsymbol{s}(t)\boldsymbol{s}(t)^\top\right]$.*

Since the source signals are independent, $\forall t \in \mathbb{Z}$ matrix $\boldsymbol{\Sigma}_{ss}(t)$ is diagonal: $\boldsymbol{\Sigma}_{ss}(t) = \mathrm{diag}(\sigma_k^2(t))$.
Then, still as in Section 3.2.1 page 11, we can *spatially* whiten the mixture signals:

- compute a matrix square root $\boldsymbol{S}$ of $\boldsymbol{\Sigma}_{xx} = \sum_t \boldsymbol{\Sigma}_{xx}(t)$;
- compute $\boldsymbol{W} = \boldsymbol{S}^{-1}$ and the whitened data $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ $\forall t \in \mathbb{Z}$.

Again, as in Section 3.2.1 we can consider a canonical BSS problem and assume $\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}\,\boldsymbol{A}^\top$, therefore matrix $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{A}$ is a rotation matrix.

Then if we consider the covariance matrices of the whitened data: $\forall t \in \mathbb{Z}$, $\boldsymbol{\Sigma}_{zz}(t) = \mathbb{E}\left[\boldsymbol{z}(t)\boldsymbol{z}(t)^\top\right]$, we get $\forall t \in \mathbb{Z}$, $\boldsymbol{\Sigma}_{zz}(t) = \boldsymbol{W}\boldsymbol{\Sigma}_{xx}(t)\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{A}\boldsymbol{\Sigma}_{ss}(t)\boldsymbol{A}^\top\boldsymbol{W}^\top = \boldsymbol{U}\boldsymbol{\Sigma}_{ss}(t)\boldsymbol{U}^\top$. Therefore, as in Section 3.3.1, matrix $\boldsymbol{U}$ can be determined by solving a joint diagonalization problem [CS96], e.g. by minimizing the following objective function:

$$J(\boldsymbol{U}) = \sum_{t} \|\boldsymbol{U}\boldsymbol{\Sigma}_{zz}(t)\boldsymbol{U}^\top - \mathrm{diag}(\boldsymbol{U}\boldsymbol{\Sigma}_{zz}(t)\boldsymbol{U}^\top)\|_F^2. \tag{1.11}$$

We thus get a variant of the SOBI algorithm [BAMCM97], summarized in Algorithm 3.

Algorithm 3 SOBI method for non-stationary sources

Estimation of the covariance matrix $\mathbf{\Sigma}_{xx}$ and diagonalization: $\mathbf{\Sigma}_{xx} = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top$ (equation (1.4))
Computation of $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}$ and of the whitening matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$ (equation (1.5))
Data whitening: $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ (equation (1.6))
Segmentation of whitened data and estimation of covariance matrices $\mathbf{\Sigma}_{zz}(t)$ on the different time frames
Approximate joint diagonalization of matrices $\mathbf{\Sigma}_{zz}(t)$ in a common basis $\boldsymbol{U}$ by minimizing (1.11)
Estimation of source signals via $\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{z}(t)$ (equation (1.7))

3.4 Time-frequency methods

So far, we have seen that:

- the use of higher order cumulants is only necessary for the non-Gaussian IID source model in Definition 1;
- second order statistics are sufficient for separating source signals that are:
  - either WSS but not IID, as in Definition 5, which amounts to exploit their *spectral* dynamics (through the autocovariance function $\boldsymbol{R}_{ss}(\tau)$);
  - or uncorrelated but not stationary, as in Definition 6, which amounts to exploit their *temporal* dynamics (through the time-varying covariance matrices $\mathbf{\Sigma}_{ss}(t)$).

The take-home message is that classical signal processing tools based on second order statistics are appropriate for performing blind separation of independent (and possibly Gaussian) sources, provided that the spectral and/or temporal source dynamics are taken into account.

However, a very simple way of highlighting the spectral and temporal dynamics of a signal is to use a *Time-frequency* (TF) representation. In this section, we will show how TF representations allow us to easily perform source separation of determined linear instantaneous mixtures. Then in Sections 4 and 5, we will see that TF representations reveal their full potential when processing convolutive and/or under-determined mixtures.

3.4.1 Time-frequency representations

Here we use the expression *time-frequency representation* to refer to complex or real-valued linear time-frequency transforms that can be implemented by means of perfect-reconstruction filterbanks [Vai93]. Classical examples of perfect reconstruction filterbanks include the *Short time Fourier transform* (STFT) (which is complex-valued) and the *Modified discrete cosine transform* (MDCT) (which is real-valued) [VVG18, chap. 2].

Every mixture signal $x_m(t)$ is thus filtered by F *analysis filters* h_f corresponding to each frequency channel $f \in \{1 \dots F\}$. The output signals are decimated by a factor $T \leq F$ to produce the F sub-band signals:

$$x_m(f, n) = (h_f * x_m)(nT), \tag{1.12}$$

where $n \in \mathbb{Z}$ is the *time frame index* and T is the *hop-size*.

Since we consider perfect-reconstruction filterbanks, we assume that there exist F *synthesis filters* g_f so that every signal $x_m(t)$ can be perfectly reconstructed from the sub-band signals: $x_m(t) = \sum_{f=1}^{F} g_f(t - nT)x_m(f, n)$.

In the same way, the source signals are decomposed in F sub-band signals $s_k(f, n) = (h_f * s_k)(nT)$ and reconstructed as

$$s_k(t) = \sum_{f=1}^{F} \sum_{n\in\mathbb{Z}} g_f(t - nT)s_k(f, n). \tag{1.13}$$

An interesting property of such a time-frequency representation is that it leaves the linear instantaneous mixture model in Definition 2 page 10 unchanged: if $\boldsymbol{x}(f, n) \in \mathbb{C}^M$ (resp. $\boldsymbol{s}(f, n) \in \mathbb{C}^K$) denotes the vector of coefficients $x_m(f, n)$ (resp. $s_k(f, n)$), then the equality $\boldsymbol{x}(t) = \boldsymbol{A}\,\boldsymbol{s}(t)\ \forall t \in \mathbb{Z}$ is equivalent to

$$\forall f \in \{1 \dots F\}, \forall n \in \mathbb{Z},\ \boldsymbol{x}(f, n) = \boldsymbol{A}\,\boldsymbol{s}(f, n). \tag{1.14}$$

Indeed, if $a_{m,k}$ denotes the entries of the mixing matrix $\boldsymbol{A}$, we have $\forall m \in \{1 \dots M\}$,

$$x_m(f,n) = (h_f * x_m)(nT) = \left(h_f * \sum_{k=1}^{K} a_{m,k} s_k\right)(nT) = \sum_{k=1}^{K} a_{m,k}(h_f * s_k)(nT) = \sum_{k=1}^{K} a_{m,k} s_k(f,n).$$

3.4.2 Time-frequency source model

Let us now introduce a general non-stationary source model. As usual we assume that the K sub-band source signals $s_k(f,n)$ are centered and independent.

In Section 3.3.1 page 14, we have presented the SOBI BSS method, that exploits the *spectral dynamics* of the source signals, by modeling them as WSS processes. Remember that the well-known *spectral representation theorem* of WSS processes [BD87] shows that the Fourier transforms of WSS processes are formed of uncorrelated random elements whose variances vary over frequency.

In a similar way, in Section 3.3.2 page 15, we have presented a variant of the SOBI method, that exploits the *temporal dynamics* of the source signals, by modeling them as sequences of uncorrelated random variables whose variances vary over time.

Now, the use of a time-frequency representation allows use to jointly exploit the spectral and the temporal dynamics, just by modeling the samples of the sub-band source signals $s_k(f,n)$ for $f \in \{1 \dots F\}$ and $n \in \mathbb{Z}$ as uncorrelated random variables whose variance $\sigma_k^2(f,n)$ depends both on the frequency channel f and the time frame n:

Definition 7 (Non-stationary TF source model) *We consider K independent source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$ and their TF representations $s_k(f,n)$ as defined in Section 3.4.1, concatenated in a vector $\boldsymbol{s}(f,n)$. For all $k \in \{1 \dots K\}$, the sub-band source signals $s_k(f,n)$ for $f \in \{1 \dots F\}$ and $n \in \mathbb{Z}$ are modeled as uncorrelated random variables of mean 0 and whose variance $\sigma_k^2(f,n)$ depends both on f and n. So $\boldsymbol{s}(f,n)$ is a random vector process of mean $\mathbb{E}[\boldsymbol{s}(f,n)] = \boldsymbol{0}$ and of TF-varying covariance matrix $\boldsymbol{\Sigma}_{ss}(f,n) = \mathbb{E}\left[\boldsymbol{s}(f,n)\boldsymbol{s}(f,n)^H\right]$.*

3.4.3 Separation method

This leads us to a new variant of the SOBI method in the determined case ($M = K$), based on the TF source model in Definition 7. Let us define the mixture covariance matrices $\boldsymbol{\Sigma}_{xx}(f,n) = \mathbb{E}[\boldsymbol{x}(f,n)\boldsymbol{x}(f,n)^H]$. Since $\boldsymbol{x}(f,n) = \boldsymbol{A}\,\boldsymbol{s}(f,n)$ (equation (1.14)), we have $\boldsymbol{\Sigma}_{xx}(f,n) = \boldsymbol{A}\boldsymbol{\Sigma}_{ss}(f,n)\boldsymbol{A}^\top$. Moreover, since the source signals are independent, matrix $\boldsymbol{\Sigma}_{ss}(f,n)$ is diagonal: $\boldsymbol{\Sigma}_{ss}(f,n) = \mathrm{diag}(\sigma_k^2(f,n))$.

Then, as in Section 3.2.1 page 11, we can *spatially* whiten the mixture signals:

- compute a matrix square root $\boldsymbol{S}$ of $\boldsymbol{\Sigma}_{xx} = \sum_{f,n} \boldsymbol{\Sigma}_{xx}(f,n)$;
- compute $\boldsymbol{W} = \boldsymbol{S}^{-1}$ and the whitened data

$$\forall f \in \{1 \dots F\}, \forall n \in \mathbb{Z},\ \boldsymbol{z}(f,n) = \boldsymbol{W}\boldsymbol{x}(f,n). \tag{1.15}$$

Again, as in Section 3.2.1 we can consider a canonical BSS problem and assume $\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}\,\boldsymbol{A}^\top$, therefore matrix $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{A}$ is a rotation matrix.

Then if we consider the covariance matrices of the whitened data: $\forall f,n$, $\boldsymbol{\Sigma}_{zz}(f,n) = \mathbb{E}\left[\boldsymbol{z}(f,n)\boldsymbol{z}(f,n)^H\right]$, we get $\boldsymbol{\Sigma}_{zz}(f,n) = \boldsymbol{W}\boldsymbol{\Sigma}_{xx}(f,n)\boldsymbol{W}^H = \boldsymbol{W}\boldsymbol{A}\boldsymbol{\Sigma}_{ss}(f,n)\boldsymbol{A}^H\boldsymbol{W}^H = \boldsymbol{U}\boldsymbol{\Sigma}_{ss}(f,n)\boldsymbol{U}^H$. Therefore, as in Section 3.3.1 page 14, matrix $\boldsymbol{U}$ can be determined by solving a joint diagonalization problem [CS96], e.g. by minimizing the following objective function:

$$J(\boldsymbol{U}) = \sum_{f,n} \|\boldsymbol{U}\boldsymbol{\Sigma}_{zz}(f,n)\boldsymbol{U}^H - \mathrm{diag}(\boldsymbol{U}\boldsymbol{\Sigma}_{zz}(f,n)\boldsymbol{U}^H)\|_F^2. \tag{1.16}$$

Finally, the source sub-band signals can be estimated as

$$\boldsymbol{y}(f,n) = \boldsymbol{U}^\top \boldsymbol{z}(f,n). \tag{1.17}$$

We thus get a variant of the SOBI algorithm [BAMCM97], summarized in Algorithm 4.

Algorithm 4 SOBI method in the TF domain

TF analysis of mixture signals: $x_m(f,n) = (h_f * x_m)(nT)$ (equation (1.12))
Estimation of the covariance matrix $\mathbf{\Sigma}_{xx}$ and diagonalization: $\mathbf{\Sigma}_{xx} = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^H$ (equation (1.4))
Computation of $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}$ and of the whitening matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$ (equation (1.5))
Data whitening: $\boldsymbol{z}(f,n) = \boldsymbol{W}\boldsymbol{x}(f,n)$ (equation (1.15))
Estimation of covariance matrices $\mathbf{\Sigma}_{zz}(f,n)$ on all time-frequency bins
Approximate joint diagonalization of matrices $\mathbf{\Sigma}_{zz}(f,n)$ in a common basis $\boldsymbol{U}$ by minimizing (1.16)
Estimation of source signals via $\boldsymbol{y}(f,n) = \boldsymbol{U}^H\boldsymbol{z}(f,n)$ (equation (1.17))
TF synthesis of source signals: $y_k(t) = \sum_{f=1}^{F}\sum_{n\in\mathbb{Z}} g_f(t-nT)y_k(f,n)$ (equation (1.13))

4 Convolutive mixtures

As already mentioned in Section 1.2 page 6, linear instantaneous mixtures cannot accurately model real acoustic mixtures, since reverberation in a room involves convolutive effects. For this reason, we now address the extension of the BSS methods presented in Section 3 page 9 to convolutive mixtures.

4.1 Source images

First, suppose that K source signals $s_k(t)$ are simultaneously emitted in a room, and that M microphones receive the observed data vector $\boldsymbol{x}(t) \in \mathbb{R}^M$. The raw source separation problem would consist in estimating the *image* of each source k, i.e. the data vector $\boldsymbol{x}_k(t) \in \mathbb{R}^M$ that would be received by the M microphones if only source k was active. These images are such that $\boldsymbol{x}(f,n) = \sum_{k=1}^{K}\boldsymbol{x}_k(f,n)$. Then the task that consists in estimating the scalar source signals $s_k(t)$ from each vector image $\boldsymbol{x}_k(t)$ is called *deconvolution* or *dereverberation*.

In this way, the source separation problem is decomposed in two steps:

- **separation**: estimate the image $\boldsymbol{x}_k(f,n)$ from the mixture $\boldsymbol{x}(f,n)$
- **deconvolution**: estimate the source signal $s_k(f,n)$ from $\boldsymbol{x}_k(f,n)$

4.2 Convolutive mixture model

Let us now introduce the convolutive mixture model in the time domain:

Definition 8 (Convolutive mixture model) *We consider K source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$, concatenated in a vector $\boldsymbol{s}(t)$. The samples of the M mixture signals $x_m(t) \in \mathbb{R}$ for $m \in \{1\dots M\}$ are then defined as*

$$x_m(t) = \sum_{k=1}^{K}(a_{mk} * s_k)(t).$$

where $\forall k \in \{1\dots K\}$, $\forall m \in \{1\dots M\}$, a_{mk} is the impulse response of a stable[3] filter. In vector form, we will write $\boldsymbol{x}(t) = \boldsymbol{A} * \boldsymbol{s}(t)$ where $\boldsymbol{A}$ is an $M \times K$ matrix of stable impulse responses a_{mk}, and $*$ denotes the convolution product between a sequence of matrices and a sequence of vectors.

The following identifiability theorem [NTJ95] generalizes Theorem 1 page 11 to the convolutive case: it proves the feasibility of ICA under mild conditions about the source signals:

Theorem 3 (Identifiability theorem) *Consider the source model in Definition 1, with the convolutive mixture model in Definition 8 in the determined case ($M = K$). Among the K IID sources s_k, suppose that at most one is Gaussian-distributed. Let $\boldsymbol{C}$ be a $K \times K$ matrix of stable impulse responses, and $\forall t \in \mathbb{Z}$, $\boldsymbol{y}(t) = \boldsymbol{C} * \boldsymbol{s}(t)$. Then the random processes $y_k(t)$ for $k \in \{1\dots K\}$ are independent if and only if matrix $\boldsymbol{C}$ is non-mixing.*

Theorem 3 shows that the source signals can be retrieved up to an unknown permutation and an unknown scale factor.

[3] Stability is defined in the *bounded-input, bounded-output* (BIBO) sense: if the input signal is bounded, then the output signal is also bounded.

4.3 Time-frequency approach

In order to simplify the problem and to make it possible to reuse the source separation methods introduced in Section 3, let us now rewrite the mixture model in Definition 8 in the time-frequency domain. More precisely, signals will be represented by their STFT, using the filterbank notation introduced in Section 3.4.1 page 16.

Moreover, we will consider the *narrow-band* approximation: we will assume that the impulse response of each mixing filter a_{mk} is short w.r.t. the time frame length of the STFT [4]. As a consequence, the spectral variations of the frequency responses $\hat{a}_{mk}(\nu)$ are slow compared to those of $\hat{h}_f(\nu)\ \forall f \in \{1 \dots F\}$. Since h_f is a very narrow band-pass filter, we will even assume that $\hat{a}_{mk}(\nu)$ is approximately constant in the pass-band of $\hat{h}_f(\nu)$ (see Figure 1.6). Consequently, we can make the approximation $\hat{h}_f(\nu)\,\hat{a}_{mk}(\nu) \approx \hat{h}_f(\nu)\,a_{mk}(f)$, where $a_{mk}(f)$ is the average value of $\hat{a}_{mk}(\nu)$ in frequency channel f. Back in the time domain, this approximation can be rewritten $(h_f * a_{mk})(t) \approx a_{mk}(f)\,h_f(t)$.

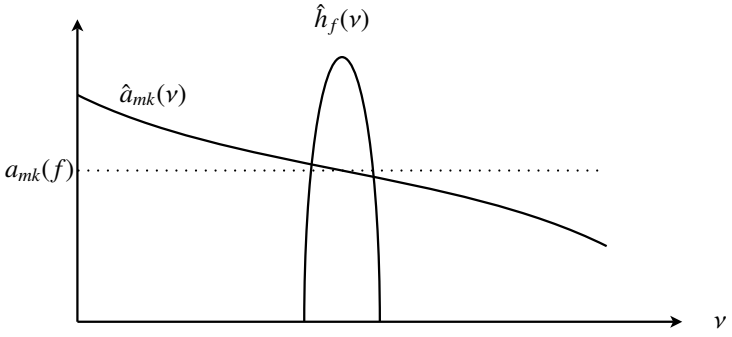


Figure 1.6: Narrow-band approximation

If we now consider the STFT of the m-th mixture signal, we get

$$\begin{aligned} x_m(f,n) &= (h_f * x_m)(nT) \\ &= \left(h_f * \left(\textstyle\sum_{k=1}^{K} a_{mk} * s_k\right)\right)(nT) \\ &= \left(\textstyle\sum_{k=1}^{K} \left(h_f * a_{mk}\right) * s_k\right)(nT) \\ &\approx \textstyle\sum_{k=1}^{K} a_{mk}(f)\left(h_f * s_k\right)(nT) \\ &= \textstyle\sum_{k=1}^{K} a_{mk}(f) s_k(f,n). \end{aligned}$$

Hence the following approximate convolutive mixture model in the time-frequency domain:

Definition 9 (TF mixture model) *We consider K source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$ and their TF representations $s_k(f,n)$ as defined in Section 3.4.1, concatenated in a vector $\boldsymbol{s}(f,n)$. For all $m \in \{1 \dots M\}$, the sub-band mixture signals $x_m(f,n)$ for $f \in \{1 \dots F\}$ and $n \in \mathbb{Z}$ are then defined as*

$$x_m(f,n) = \sum_{k=1}^{K} a_{mk}(f) s_k(f,n),$$

or in matrix form,

$$\boldsymbol{x}(f,n) = \boldsymbol{A}(f)\boldsymbol{s}(f,n), \tag{1.18}$$

where $\boldsymbol{A}(f) \in \mathbb{C}^{M \times K}$ is the matrix of entries a_{mk}.

It can be noted that in each frequency channel f, (1.18) is a linear instantaneous mixture model (to be compared to equation (1.2) page 10), parameterized by the mixing matrix $\boldsymbol{A}(f)$. Therefore we are tempted to apply any ICA method designed for the linear instantaneous mixture model, such as those described in Section 3, in every frequency channel of the STFT, in order to estimate the sub-band signals $s_k(f,n)$.

[4] Note that this assumption is not realistic: the length of the impulse response a_{mk} corresponds to the reverberation time, which is usually several hundreds of miliseconds, while the typical time frame length in an STFT is a few tens of miliseconds. Nevertheless, this approximation is often used in audio source separation methods because it leads to a very simple mixture model in the TF domain (*cf.* Definition 9), which proves to perform well in various applications [OF10].

4.4 Independent component analysis

As in Section 3.1.2 page 10, it may seem natural to estimate the sub-band source signals as linear instantaneous combinations of the sub-band mixture signals: $\boldsymbol{y}(f,n) = \boldsymbol{B}(f)\,\boldsymbol{x}(f,n)$, where the entries of vector $\boldsymbol{y}(f,n)$ are the sub-band source signal estimates, and $\boldsymbol{B}(f) \in \mathbb{C}^{K\times M}$ is referred to as the *separation matrix*. Then the source separation problem amounts to finding an optimal separation matrix.

Linear source separation is generally feasible in the case of determined and over-determined mixtures:

- if $M = K$ and if matrix $\boldsymbol{A}(f)$ is invertible, then the separation matrix $\boldsymbol{B}(f) = \boldsymbol{A}(f)^{-1}$ leads to $\boldsymbol{y}(f,n) = \boldsymbol{s}(f,n)$;
- more generally, if $M \geq K$ and if matrix $\boldsymbol{A}(f)$ has full rank, then the separation matrix $\boldsymbol{B}(f) = \boldsymbol{A}(f)^{\dagger}$ leads to $\boldsymbol{y}(f,n) = \boldsymbol{s}(f,n)$.

However, in the under-determined case ($M < K$), linear source separation is generally not feasible (*cf.* Section 5).

In the determined case ($M = K$), independent component analysis [CJ10] can be applied in each frequency channel f. It aims to find a *separation matrix* $\boldsymbol{B}(f)$ that makes the K sub-band signals $y_k(f,n)$ independent. Then, as in Section 3.2 page 11, we get $\boldsymbol{y}(f,n) = \boldsymbol{C}(f)\boldsymbol{s}(f,n)$, where $\boldsymbol{C}(f) = \boldsymbol{B}(f)\boldsymbol{A}(f)$ is a non-mixing matrix (*cf.* Definition 3 page 10).

4.5 Indeterminacies

While the indeterminacies induced by the non-mixing matrix $\boldsymbol{C}$ were acceptable when we were considering linear instantaneous mixtures in Section 3, we now encounter an unexpected issue, because with the TF mixture model in Definition 9, there are F possibly different non-mixing matrices $\boldsymbol{C}(f)$. The problem is that, for instance, the permutations can be different in two different frequency channels f. If we choose to ignore that, and to just reconstruct the source signals $y_k(t)$ from the separated sub-band signals $y_k(f,n)$ with the synthesis filters g_f, it is very likely that the resulting signals $y_k(t)$ will be formed of different sources s_k in different frequency channels. In other words, reconstructing the source signals from the separated sub-band signals would amount to remix the estimated sources!

In order to avoid this problem, we need to solve the permutation indeterminacy in the frequency channels of the STFT. Note that this multiple permutation issue is inherent to the time-frequency approach and to the narrow-band approximation introduced in Section 4.3: if BSS was performed in the time domain instead, Theorem 3 page 18 proves that all sources can theoretically be retrieved up to a unique permutation.

Even though, assuming that we do have a method that allows us to solve the multiple permutation indeterminacy, the other indeterminacy remains: there is an unknown multiplicative factor associated to each source in each frequency channel f.

Actually, both kinds of indeterminacies can be solved jointly, by introducing additional assumptions about the mixing filters a_{mk} and/or the source signals s_k:

- regarding the source signals, we can assume that the temporal dynamics (over n) of $\sigma_k^2(f,n)$ are similar between different frequency channels f for a same source k (nonparametric approach), or we can also exploit a parametric model, such as the *Nonnegative Matrix Factorization (NMF)* [VVG18, chap. 8] [OF10].
- regarding the mixing filters, we can assume that their frequency responses $a_{mk}(f)$ are slowly varying w.r.t. f (nonparametric approach), or we can also exploit a parametric mixture model, such as the beamforming model or the anechoic model. The beamforming model [VVG18, chap. 10] relies on the plane wave and far field hypotheses (no reverberation) and assumes that the microphone antenna is linear. In this case, we get $a_{mk}(f) = e^{-2\imath\pi f\tau_{mk}}$ where $\tau_{mk} = \frac{d_m}{c}\sin(\theta_k)$, where parameters d_m denote the positions of the sensors on the linear antenna and parameters θ_k denote the angles of the sources (see Figure 1.7). The anechoic model is a bit more general: it assumes that the sources are punctual and that there is no reverberation. In this case, we get $a_{mk}(f) = \alpha_{mk}e^{-2\imath\pi f\tau_{mk}}$ where $\alpha_{mk} = \frac{1}{\sqrt{4\pi}r_{mk}}$, $\tau_{mk} = \frac{r_{mk}}{c}$, and parameters r_{mk} denote the distances between the sensors and sources. In practice, none of these two mixture models is able to accurately represent real acoustic mixtures; nevertheless they can be helpful to solve the multiple permutation problem.

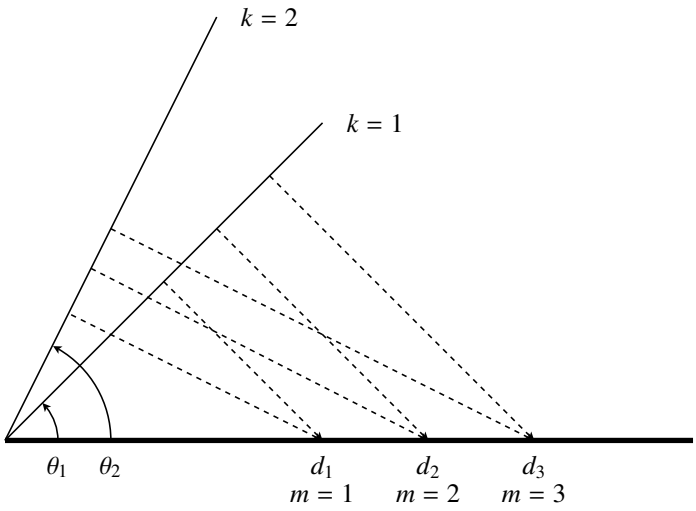


Figure 1.7: Beamforming mixture model

5 Under-determined mixtures

As mentioned in Section 1.1 page 5, linear source separation is generally not feasible in the under-determined case, because there are infinitely many solutions. Without additional information about the mixture or the source signals, it is impossible to retrieve the original sources from the mixture signals.

Unfortunately, the under-determined case is often encountered in audio signal processing: indeed, many audio signals are either monophonic ($M = 1$) or stereophonic ($M = 2$), whereas the number of sources K is generally greater than 2. In this section, we will see how additional information can be taken into account to perform source separation in such a challenging scenario.

5.1 Under-determined convolutive mixtures

We still consider the TF mixture model in Definition 9 page 19 : $\boldsymbol{x}(f,n) = \boldsymbol{A}(f)\boldsymbol{s}(f,n)$ (equation (1.18)), and the TF source model in Definition 7 page 17: the samples of the sub-band source signals $s_k(f,n)$ for $f \in \{1 \dots F\}$ and $n \in \mathbb{Z}$ are uncorrelated random variables whose variance $\sigma_k^2(f,n)$ depends both on the frequency channel f and the time frame n, so that the covariance matrix of $\boldsymbol{s}(f,n)$ is $\boldsymbol{\Sigma}_{ss}(f,n) = \mathrm{diag}(\sigma_k^2(f,n))$.

As in Section 4.4 page 20, it may seem natural to estimate the sub-band source signals as linear instantaneous combinations of the sub-band mixture signals: $\boldsymbol{y}(f,n) = \boldsymbol{B}(f)\,\boldsymbol{x}(f,n)$, where the entries of vector $\boldsymbol{y}(f,n)$ are the sub-band source signal estimates, and $\boldsymbol{B}(f) \in \mathbb{C}^{K\times M}$ is referred to as the *separation matrix*. Then the source separation problem amounts to finding an optimal separation matrix.

However if $M < K$, even if the mixing matrix $\boldsymbol{A}(f)$ and the source model $\boldsymbol{\Sigma}_{ss}(f,n)$ were known, the exact separation would not be feasible: there is no matrix $\boldsymbol{B}(f)$ such that $\boldsymbol{B}(f)\,\boldsymbol{A}(f) = \boldsymbol{I}_K$, because the maximum rank of matrix $\boldsymbol{B}(f)\,\boldsymbol{A}(f)$ is $M < K$ (*cf.* Figure 1.8).

However, we can still try to find an approximate solution $\boldsymbol{y}(f,n)$ in the least squares sense.

5.2 Separation via non-stationary filtering

First, we suppose that the mixing matrix $\boldsymbol{A}(f)$ and the source model $\boldsymbol{\Sigma}_{ss}(f,n)$ are known. Even though only an approximate solution $\boldsymbol{y}(f,n)$ can be obtained, this solution can be improved by adding a degree of freedom to the separation matrix $\boldsymbol{B}$: we will now make it depend on time n, so that the estimate $\boldsymbol{y}(f,n)$ is obtained by *non-stationary filtering*:

$$\boldsymbol{y}(f,n) = \boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) \tag{1.19}$$

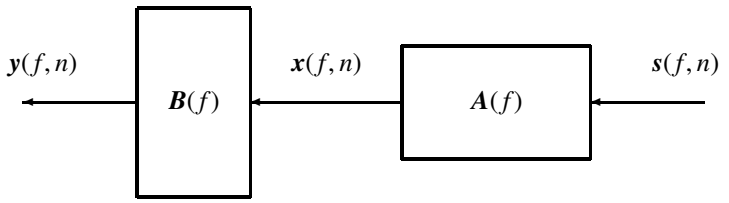


Figure 1.8: Under-determined mixtures: there is no matrix $\boldsymbol{B}(f)$ such that $\boldsymbol{B}(f)\,\boldsymbol{A}(f) = \boldsymbol{I}_K$

where $\boldsymbol{B}(f,n) \in \mathbb{C}^{K\times M}$. The approximate solution will be found in the least squares sense, by considering the *Minimum mean square error* (MMSE) estimator [SAK+13]: we will look for the separation matrix $\boldsymbol{B}(f,n)$ which minimizes the *Mean square error* (MSE) $\mathbb{E}[\|\boldsymbol{y}(f,n) - \boldsymbol{s}(f,n)\|_2^2]$. This MSE is such that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{y}(f,n) - \boldsymbol{s}(f,n)\|_2^2] &= \mathbb{E}\left[(\boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) - \boldsymbol{s}(f,n))^H\,(\boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) - \boldsymbol{s}(f,n))\right] \\
&= \text{trace}\left(\mathbb{E}\left[(\boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) - \boldsymbol{s}(f,n))\,(\boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) - \boldsymbol{s}(f,n))^H\right]\right) \\
&= \text{trace}\left(\boldsymbol{B}(f,n)\boldsymbol{\Sigma}_{xx}(f,n)\boldsymbol{B}(f,n)^H - \boldsymbol{B}(f,n)\boldsymbol{\Sigma}_{xs}(f,n) - \boldsymbol{\Sigma}_{sx}(f,n)\boldsymbol{B}(f,n)^H + \boldsymbol{\Sigma}_{ss}(f,n)\right).
\end{aligned}$$

The MSE is minimized when the Wirtinger matrix gradient [GR65] w.r.t. $\boldsymbol{B}(f,n)$ is zero:

$$\boldsymbol{B}(f,n)\boldsymbol{\Sigma}_{xx}(f,n) - \boldsymbol{\Sigma}_{sx}(f,n) = \boldsymbol{0}.$$

Therefore the solution is given by $\boldsymbol{B}(f,n) = \boldsymbol{\Sigma}_{sx}(f,n)\,\boldsymbol{\Sigma}_{xx}(f,n)^{-1}$, where $\boldsymbol{\Sigma}_{xx}(f,n) = \boldsymbol{A}(f)\,\boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H$ and $\boldsymbol{\Sigma}_{sx}(f,n) = \boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H$. We can finally express the MMSE estimator as (1.19), with

$$\boldsymbol{B}(f,n) = \boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H\left(\boldsymbol{A}(f)\,\boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H\right)^{-1}. \tag{1.20}$$

We remark that the MMSE estimator guarantees the perfect reconstruction of the mixture signals from the estimated source signals: $\boldsymbol{A}(f)\,\boldsymbol{y}(f,n) = \boldsymbol{x}(f,n)$.

This MMSE estimator is also known as the *generalized* or *multichannel* Wiener filter, because in the particular case of monophonic mixtures ($M = 1$), it boils down to the well-known Wiener filter. Indeed, because of the scale indeterminacy of the model, we can assume without loss of generality that $\boldsymbol{A}(f) = [1,\ldots,1]$. Then the MMSE estimator defined by (1.19) and (1.20) can be rewritten as $y_k(f,n) = \frac{\sigma_k^2(f,n)}{\sum_{l=1}^K \sigma_l^2(f,n)}\,x(f,n)$, which is the usual form of the Wiener filter.

In practice of course, the mixing matrix $\boldsymbol{A}(f)$ and the source model $\boldsymbol{\Sigma}_{ss}(f,n)$ are unknown; they thus have to be estimated from the observed data. For instance, $\boldsymbol{A}(f)$ can be assumed slowly varying over f (nonparametric approach), or parameterized according to the beamforming or the anechoic model introduced in Section 4.5 page 20, and $\boldsymbol{\Sigma}_{ss}(f,n)$ can be assumed sparse in the TF domain as in Section 5.3.1 (nonparametric approach), or parameterized according to an NMF model [VVG18, chap. 8] [OF10].

The resulting algorithm is sketched in Algorithm 5.

Algorithm 5 Under-determined source separation in the TF domain

TF analysis of mixture signals: $x_k(f,n) = (h_f * x_k)(nT)$ (equation (1.12))
Estimation of $\boldsymbol{A}(f)$ and $\sigma_k^2(f,n)$
Computation of $\boldsymbol{B}(f,n) = \boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H\left(\boldsymbol{A}(f)\,\boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H\right)^{-1}$ (equation (1.20))
Estimation of source sub-band signals as $\boldsymbol{y}(f,n) = \boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n)$ (equation (1.19))
TF synthesis of source signals: $y_k(t) = \sum_{f=1}^F \sum_{n\in\mathbb{Z}} g_f(t-nT)y_k(f,n)$ (equation (1.13))

5.3 Stereophonic mixtures: separation based on sparsity

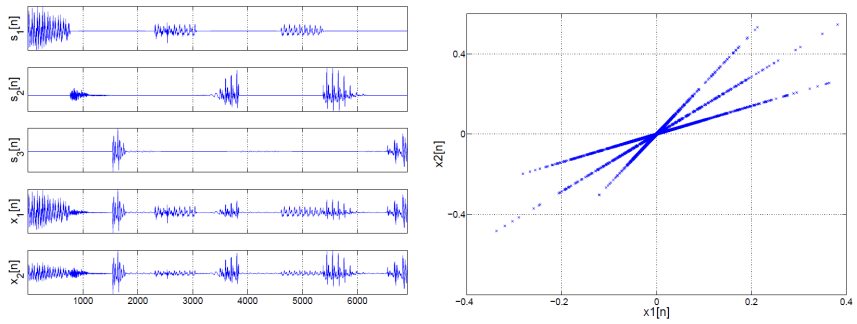
5.3.1 Temporal sparsity

We now consider the particular case of stereophonic ($M = 2$) linear instantaneous mixtures as in Definition 2 page 10 (defined by a unique mixing matrix $\boldsymbol{A}$ in order to simplify, but this approach would also work with the TF mixture model in Definition 9 page 19), so that the mixture model in the time domain is $\boldsymbol{x}(t) = \boldsymbol{A}\,\boldsymbol{s}(t)$.

We consider the example in Figure 1.9-(a): the $K = 3$ source signals represented in the three top lines of the figure are never active at the same time. We say that they are *sparse* in the time domain, in the sense that most of their temporal samples are zero. The $M = 2$ mixture signals are represented in the two bottom lines. Clearly, in this particular case of non-overlapping source signals, the source separation problem is a simple *classification* problem: it amounts to *segment* the mixture signals, and label the successive segments as "source 1", "source 2", etc.

In this simple scenario, the classification of the time samples can be easily performed by plotting the *dispersion diagram*, represented in Figure 1.9-(b): for every time t, the point of coordinates $(x_1(t), x_2(t))$ is drawn in the plane. This diagram clearly makes appear three straight lines, which correspond to the three sources. Indeed, when only source k is active, the linear instantaneous mixture model in Definition 2 yields $\boldsymbol{x}(t) = \boldsymbol{a}_k\, s_k(t)$, where $\boldsymbol{a}_k$ is the k-th column of matrix $\boldsymbol{A}$ (remember that because of the scale indeterminacy, we can assume without loss of generality that $\boldsymbol{a}_k$ is a unit vector). Therefore all points of coordinates $\boldsymbol{x}(t)$ in the plane that are generated by source k belong to the straight line passing through the origin and defined by the direction vector $\boldsymbol{a}_k$, and their position on this straight line corresponds to the value of the time sample $s_k(t)$.

Therefore in this simple case, source separation can be very easily performed by detecting the lines in the dispersion diagram (e.g. by using the Hough transform [SS01]), which makes it possible to jointly estimate the column vectors $\boldsymbol{a}_k$ and the number of sources K. Then the images of the sources (defined in Section 4.1 page 18) can be retrieved by selecting the points $\boldsymbol{x}(t)$ that are the closest to each line, and finally the source signals $s_k(t)$ can be estimated by calculating the positions of these points on the line.



(a) Signaux sources temporels et mélange stéréo correspondant

(b) Diagrammes de dispersion (x_1, x_2) en temps.

Figure 1.9: Sparsity in time domain

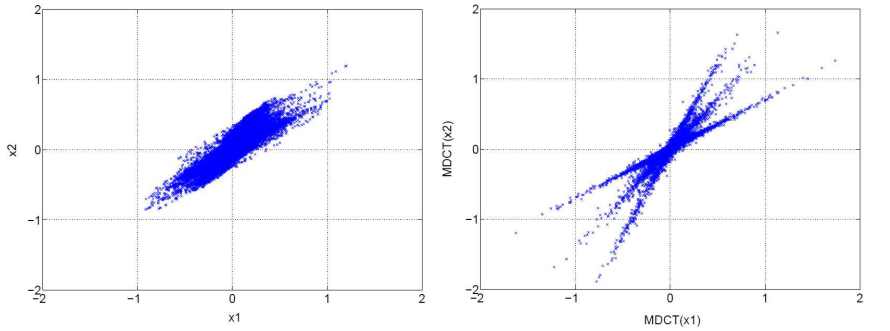
5.3.2 Sparsity in a transformed domain

Unfortunately, the example in Figure 1.9 is not very realistic: it rarely happens, especially in music, that the different source signals do not overlap in the time domain. Figure 1.10-(a) shows an example of dispersion diagram obtained with a mixture of overlapping music sources: no straight line emerges from the cloud of points.

However, even when they do overlap in time, audio signals are often sparse in the TF domain: the spectrum of various sounds (especially in music) is made of a discrete set of frequencies. Figure 1.10-(b) shows another dispersion diagram obtained from the same mixture of music sources as in Figure 1.10-(a), except that the coordinates of the points are not obtained from the time samples $\boldsymbol{x}(t)$, but from the MDCT TF transform[5] $\boldsymbol{x}(f, n)$. The resulting

[5]The MDCT is known to produce very sparse TF representations, which is why it is widely used in lossy audio data compression [Luo09].

dispersion diagram is not as clean as that of Figure 1.9-(b), but again three straight lines clearly emerge from the cloud of points, which shows that the mixture is made of $K = 3$ sources, which can be separated in the same way as in Section 5.3.1, but in the TF domain.



(a) Échantillons temporels.

(b) Coefficients temps-frequence après décomposition par MDCT.

Figure 1.10: Sparsity in TF domain

5.3.3 DUET method

We can now introduce the celebrated *Degenerate unmixing estimation technique* (DUET) method [JRY00, Ric07], which is dedicated to stereophonic ($M = 2$) mixtures, in the linear instantaneous mixture case: $\boldsymbol{x}(f,n) = \boldsymbol{A}\,\boldsymbol{s}(f,n)$ (equation 1.14 page 16). Without loss of generality, the column vectors of the mixing matrix $\boldsymbol{A}$ are parameterized as $\boldsymbol{a}_k = \begin{bmatrix} \cos(\theta_k) \\ \sin(\theta_k) \end{bmatrix}$, where $\theta_k \in \mathbb{R}$. Regarding the source signals, we consider a *sparse* source model in the TF domain:

Definition 10 (Sparse TF source model) *We consider the same source model as in Definition 7. In addition, we assume that* $\forall f,n$, *there is a unique* $k_{(f,n)} \in \{1 \dots K\}$ *such that* $\sigma^2_{k_{(f,n)}}(f,n) > 0$, *and* $\forall l \neq k_{(f,n)}$, $\sigma^2_l(f,n) = 0$.

This additional assumption means that only one source can be active at any TF bin (f,n) (sources do not overlap in the TF domain). Therefore $\forall f,n$, $\boldsymbol{x}(f,n) = \boldsymbol{a}_{k_{(f,n)}} s_{k_{(f,n)}}(f,n)$.

The DUET method then consists of two steps: parameter estimation and source separation.

In the first step, the TF representation of the mixture signals is first computed by using the analysis filters h_f as in equation (1.12) page 16. Then, in order to estimate the mixture parameters θ_k, the histogram of the angles of vectors $\boldsymbol{x}(f,n)$ is first computed. The peaks of this histogram theoretically correspond to the angles θ_k, which can thus be estimated by performing peak detection. Then the active source $k_{(f,n)}$ at frequency bin (f,n) is estimated by selecting the angle θ_k which is the closest to the angle of the observed vector $\boldsymbol{x}(f,n)$.

In the second step, the source images (defined in Section 4.1 page 18) are estimated via binary masking [YR04]:

$$\forall k \in \{1 \dots K\},\ \boldsymbol{y}_k(f,n) = \begin{cases} \boldsymbol{x}(f,n)\ \forall (f,n) \text{ such that } k_{(f,n)} = k, \\ \boldsymbol{0} \text{ for the other time-frequency bins } (f,n). \end{cases} \tag{1.21}$$

Then the sub-band source signals are estimated with the MMSE estimator introduced in equation (1.20) page 22, which here boils down to a zero separation matrix $\boldsymbol{B}(f,n)$, except its k-th row which is[6]

$$\boldsymbol{a}_k(f)^\dagger = \frac{\boldsymbol{a}_k(f)^H}{\|\boldsymbol{a}_k(f)\|_2^2}. \tag{1.22}$$

Therefore the estimate of the k-th sub-band source signal as defined in (1.19) page 21 is

$$y_k(f,n) = \boldsymbol{a}_k(f)^\dagger \boldsymbol{y}_k(f,n). \tag{1.23}$$

[6] In equation (1.20), matrix $\boldsymbol{\Sigma}_{ss}$ is singular, so the matrix inverse is replaced by the matrix pseudo-inverse, leading to equation (1.22).

Finally, the source signals are reconstructed in the time domain by using the synthesis filters g_f, as in equation (1.13) page 16. The DUET method is summarized in Algorithm 6.

Algorithm 6 DUET method

TF analysis of mixture signals: $x_k(f,n) = (h_f * x_k)(nT)$ (equation (1.12))
Estimation of parameters θ_k and of the active source $k_{(f,n)}$
    Computation of the histogram of the angles of vectors $\boldsymbol{x}(f,n)$
    Peak detection in order to estimate parameters θ_k
    Determination of the active source at (f,n) by proximity with θ_k
Source separation:
    Estimation of source images $\boldsymbol{y}_k(f,n)$ via binary masking (equation (1.21))
    MMSE estimation of sub-band source signals: $y_k(f,n) = \boldsymbol{a}_k(f)^\dagger \boldsymbol{y}_k(f,n)$ (equation (1.23))
TF synthesis of source signals: $y_k(t) = \sum_{f=1}^{F} \sum_{n\in\mathbb{Z}} g_f(t-nT) y_k(f,n)$ (equation (1.13))

6 Conclusion

In this chapter, we have reviewed several source separation models and methods, dedicated to determined linear instantaneous mixtures, determined convolutive mixtures, and under-determined mixtures. All these methods exploit the spatial diversity of the observed mixture signals (they require that $M > 1$), and their funding principle is that all source signals are statistically independent.

Source separation requires to make assumptions about the mixture and about the source signals, which are generally expressed in terms of probability distributions. For (over-)determined linear mixtures, we have seen that assuming independent sources is generally sufficient to make the separation possible (under mild conditions on the source probability distributions). When the mixture is under-determined however, it is necessary to make additional assumptions about the mixture and about the source signals, that can be formulated either in a non-parametric way (via regularization), or by exploiting parametric models.

This chapter forms an introduction to audio source separation, with a selection of models and methods; several topics that have been investigated in the literature could not been addressed here, for instance:

- The separation of non-stationary mixtures requires to develop adaptive separation algorithms [CL96];
- *Informed source separation* techniques exploit some possibly available extra information about the mixture or the sources, such as the spatial positions of the sources and microphones (e.g. via *beamforming*), or the transcription of the source signals (speech or music) [EM12];
- Deep learning techniques are able to automatically learn how to perform separation from a large database of source and mixture signals [VVG18].
- Criteria for the objective assessment of audio source separation are required in order to compare the performance of various separation methods [VGF06].

Bibliography

[BAMCM97] Adel Belouchrani, Karim Abed-Meraim, Jean-François Cardoso, and Eric Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[BD87] Peter J. Brockwell and Richard A. Davis. *The Spectral Representation of a Stationary Process*, pages 112–158. Springer New York, New York, NY, 1987.

[Car98] Jean-François Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[CJ10] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, Inc. (Elsevier), USA, 1st edition, 2010.

[CL96] Jean-François Cardoso and B. H. Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, 1996.

[Com94] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287 – 314, April 1994. Special issue on Higher-Order Statistics.

[CS93] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, 1993.

[CS96] Jean-François Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, 1996.

[EM12] Sebastian Ewert and Meinard Müller. *Multimodal Music Processing*, volume 3, chapter Score-Informed Source Separation for Music Signals, pages 73–94. January 2012.

[GR65] Robert C. Gunning and Hugo Rossi. *Analytic Functions of Several Complex Variables*. AMS Chelsea Publishing. Prentice-Hall, Englewood Cliffs, N.J., USA, 1965.

[JRY00] A. Jourjine, Scott Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In *Proc. of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2985–2988, 2000.

[Luo09] Fa-Long Luo. *Mobile Multimedia Broadcasting Standards: Technology and Practice*. Springer Science & Business Media, January 2009.

[NTJ95] Hoang-Lan Nguyen Thi and Christian Jutten. Blind source separation for convolutive mixtures. *Signal Processing*, 45(2):209 – 229, 1995.

[OF10] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010.

[Ric07] Scott Rickard. *The DUET Blind Source Separation Algorithm*, pages 217–241. Springer, Dordrecht, Netherlands, 2007.

[SAK+13] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and Hiroshi Sawada. A multichannel MMSE-based framework for speech source separation and noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1913–1928, 2013.

[SS01] George Stockman and Linda G. Shapiro. *Computer Vision*. Prentice Hall PTR, USA, 1st edition, 2001.

[Vai93] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice-Hall, Inc., USA, 1993.

[VGF06] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

[VVG18] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot. *Audio Source Separation and Speech Enhancement*. Wiley Publishing, 1st edition, 2018.

[YR04] O. Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.