Audio source separation

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

TSIA 206 - Speech and audio processing

Part I

Introduction

Introduction

- Source separation
  - Art of estimating "source" signals, assumed independent, from the observation of one or several "mixtures" of these sources
- Application examples:
  - Denoising (cocktail party, suppression of vuvuzela, karaoke)
  - Separation of the instruments in polyphonic music
  - Remix, transformations, re-spatialization

Introduction

- Source separation
  - Art of estimating "source" signals, assumed independent, from the observation of one or several "mixtures" of these sources
- Application examples:
  - Denoising (cocktail party, suppression of vuvuzela, karaoke)
  - Separation of the instruments in polyphonic music
  - Remix, transformations, re-spatialization

Introduction

- Source separation
  - Art of estimating "source" signals, assumed independent, from the observation of one or several "mixtures" of these sources
- Application examples:
  - Denoising (cocktail party, suppression of vuvuzela, karaoke)
  - Separation of the instruments in polyphonic music
  - Remix, transformations, re-spatialization

Introduction

- Source separation
  - Art of estimating "source" signals, assumed independent, from the observation of one or several "mixtures" of these sources
- Application examples:
  - Denoising (cocktail party, suppression of vuvuzela, karaoke)
  - Separation of the instruments in polyphonic music
  - Remix, transformations, re-spatialization

Introduction

- Source separation
  - Art of estimating "source" signals, assumed independent, from the observation of one or several "mixtures" of these sources
- Application examples:
  - Denoising (cocktail party, suppression of vuvuzela, karaoke)
  - Separation of the instruments in polyphonic music
  - Remix, transformations, re-spatialization

Typology of the mixture models

- Definition of the problem
  - Observations: M mixtures $x_m(t)$, concatenated in a vector $\mathbf{x}(t)$
  - Unknowns: K sources $s_k(t)$, concatenated in a vector $\mathbf{s}(t)$
  - General mixture model: function $\mathscr{A}$ which transforms $\mathbf{s}(t)$ into $\mathbf{x}(t)$
- Stationarity: $\mathscr{A}$ is translation invariant
- Linearity: $\mathscr{A}$ is a linear map
- Memory:
  - Convolutive mixtures
  - Instantaneous mixtures: $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$
    - $\mathscr{A}$ is defined by the "mixture matrix" $\mathbf{A}$ (of dimension $M \times K$)
- Inversibility:
  - Determined mixtures: $M = K$
  - Over-determined mixtures: $M > K$
  - Under-determined mixtures: $M < K$

Typology of the mixture models

- Definition of the problem
  - Observations: M mixtures $x_m(t)$, concatenated in a vector $\mathbf{x}(t)$
  - Unknowns: K sources $s_k(t)$, concatenated in a vector $\mathbf{s}(t)$
  - General mixture model: function $\mathscr{A}$ which transforms $\mathbf{s}(t)$ into $\mathbf{x}(t)$
- Stationarity: $\mathscr{A}$ is translation invariant
- Linearity: $\mathscr{A}$ is a linear map
- Memory:
  - Convolutive mixtures
  - Instantaneous mixtures: $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$
    - $\mathscr{A}$ is defined by the "mixture matrix" $\mathbf{A}$ (of dimension $M \times K$)
- Inversibility:
  - Determined mixtures: $M = K$
  - Over-determined mixtures: $M > K$
  - Under-determined mixtures: $M < K$

Typology of the mixture models

- Definition of the problem
  - Observations: M mixtures $x_m(t)$, concatenated in a vector $\mathbf{x}(t)$
  - Unknowns: K sources $s_k(t)$, concatenated in a vector $\mathbf{s}(t)$
  - General mixture model: function $\mathscr{A}$ which transforms $\mathbf{s}(t)$ into $\mathbf{x}(t)$
- Stationarity: $\mathscr{A}$ is translation invariant
- Linearity: $\mathscr{A}$ is a linear map
- Memory:
  - Convolutive mixtures
  - Instantaneous mixtures: $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$
    - $\mathscr{A}$ is defined by the "mixture matrix" $\mathbf{A}$ (of dimension $M \times K$)
- Inversibility:
  - Determined mixtures: $M = K$
  - Over-determined mixtures: $M > K$
  - Under-determined mixtures: $M < K$

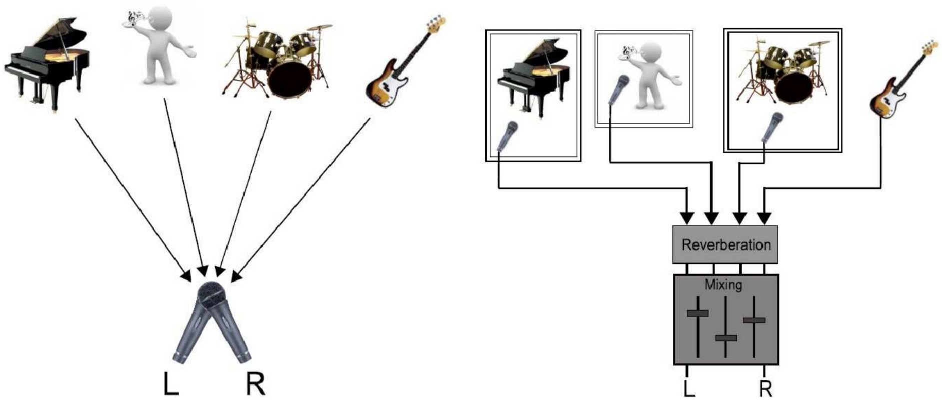
Typology of the mixture models

- Definition of the problem
  - Observations: M mixtures $x_m(t)$, concatenated in a vector $\mathbf{x}(t)$
  - Unknowns: K sources $s_k(t)$, concatenated in a vector $\mathbf{s}(t)$
  - General mixture model: function $\mathscr{A}$ which transforms $\mathbf{s}(t)$ into $\mathbf{x}(t)$
- Stationarity: $\mathscr{A}$ is translation invariant
- Linearity: $\mathscr{A}$ is a linear map
- Memory:
  - Convolutive mixtures
  - Instantaneous mixtures: $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$
    - $\mathscr{A}$ is defined by the "mixture matrix" $\mathbf{A}$ (of dimension $M \times K$)
- Inversibility:
  - Determined mixtures: $M = K$
  - Over-determined mixtures: $M > K$
  - Under-determined mixtures: $M < K$

Typology of the mixture models

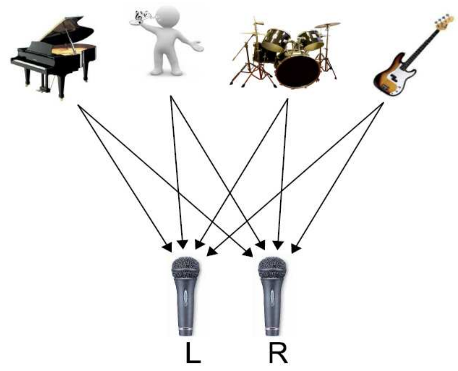
- Definition of the problem
  - Observations: M mixtures $x_m(t)$, concatenated in a vector $\mathbf{x}(t)$
  - Unknowns: K sources $s_k(t)$, concatenated in a vector $\mathbf{s}(t)$
  - General mixture model: function $\mathscr{A}$ which transforms $\mathbf{s}(t)$ into $\mathbf{x}(t)$
- Stationarity: $\mathscr{A}$ is translation invariant
- Linearity: $\mathscr{A}$ is a linear map
- Memory:
  - Convolutive mixtures
  - Instantaneous mixtures: $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$
    - $\mathscr{A}$ is defined by the "mixture matrix" $\mathbf{A}$ (of dimension $M \times K$)
- Inversibility:
  - Determined mixtures: $M = K$
  - Over-determined mixtures: $M > K$
  - Under-determined mixtures: $M < K$

Instantaneous linear mixtures

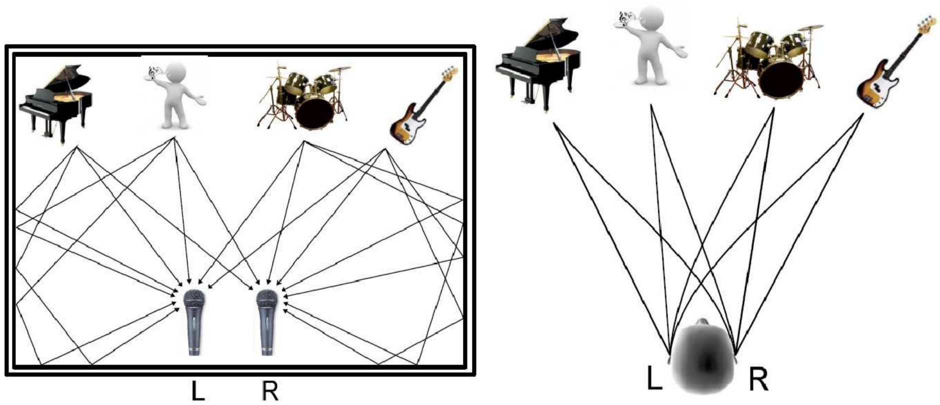


(a) XY Stereo configuration (b) Direct injection to the mixer

Anechoic linear mixtures



Convolutive linear mixtures



(a) Convolutive mixture (b) Binaural mixture

Part II

Mathematical reminders

Real random vectors

- Notation: $\phi[\mathbf{x}]$ denotes a function of $p(\mathbf{x})$
- Mean vector: $\mu_x = \mathbb{E}[\mathbf{x}]$
- Covariance matrix: $\Sigma_{xx} = \mathbb{E}[(\mathbf{x}-\mu_x)(\mathbf{x}-\mu_x)^T]$
- Characteristic function: $\phi_x(\mathbf{f}) = \mathbb{E}[e^{-2i\pi\mathbf{f}^T\mathbf{x}}] = \int_{\mathbb{R}} p(\mathbf{x}) e^{-2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{x}$
- Probability distribution: $p(\mathbf{x}) = \int_{\mathbb{R}} \phi_x(\mathbf{f}) e^{+2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{f}$
- Cumulants:
  - Definition: $\ln(\phi_x(\mathbf{f})) = \sum_{n=1}^{+\infty} \frac{(-2i\pi)^n}{n!} \sum_{k_1=1}^{K} \sum_{k_n=1}^{K} \kappa^n_{k_1\dots k_n}[\mathbf{x}] f_{k_1}\dots f_{k_n}$
  - $\kappa^n[\mathbf{x}]$ is an n-th order tensor
  - $\kappa^1[\mathbf{x}]$ is the mean vector, $\kappa^2[\mathbf{x}]$ is the covariance matrix
  - If $p(\mathbf{x})$ is symmetric ($p(-\mathbf{x}) = p(\mathbf{x})$), $\kappa^n[\mathbf{x}] = 0$ for any odd value n
  - the ratio $\kappa^4_{k,k,k,k}[\mathbf{x}]/(\kappa^2_{k,k}[\mathbf{x}])^2$ is called "kurtosis"

Real random vectors

- Notation: $\phi[\mathbf{x}]$ denotes a function of $p(\mathbf{x})$
- Mean vector: $\mu_x = \mathbb{E}[\mathbf{x}]$
- Covariance matrix: $\Sigma_{xx} = \mathbb{E}[(\mathbf{x}-\mu_x)(\mathbf{x}-\mu_x)^T]$
- Characteristic function: $\phi_x(\mathbf{f}) = \mathbb{E}[e^{-2i\pi\mathbf{f}^T\mathbf{x}}] = \int_{\mathbb{R}} p(\mathbf{x}) e^{-2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{x}$
- Probability distribution: $p(\mathbf{x}) = \int_{\mathbb{R}} \phi_x(\mathbf{f}) e^{+2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{f}$
- Cumulants:
  - Definition: $\ln(\phi_x(\mathbf{f})) = \sum_{n=1}^{+\infty} \frac{(-2i\pi)^n}{n!} \sum_{k_1=1}^{K} \sum_{k_n=1}^{K} \kappa^n_{k_1\dots k_n}[\mathbf{x}] f_{k_1}\dots f_{k_n}$
  - $\kappa^n[\mathbf{x}]$ is an n-th order tensor
  - $\kappa^1[\mathbf{x}]$ is the mean vector, $\kappa^2[\mathbf{x}]$ is the covariance matrix
  - If $p(\mathbf{x})$ is symmetric ($p(-\mathbf{x}) = p(\mathbf{x})$), $\kappa^n[\mathbf{x}] = 0$ for any odd value n
  - the ratio $\kappa^4_{k,k,k,k}[\mathbf{x}]/(\kappa^2_{k,k}[\mathbf{x}])^2$ is called "kurtosis"

9/48 Une école de l'IMT Audio source separation

Real random vectors

- Notation: $\phi[\mathbf{x}]$ denotes a function of $p(\mathbf{x})$
- Mean vector: $\mu_x = \mathbb{E}[\mathbf{x}]$
- Covariance matrix: $\Sigma_{xx} = \mathbb{E}[(\mathbf{x}-\mu_x)(\mathbf{x}-\mu_x)^T]$
- Characteristic function: $\phi_x(\mathbf{f}) = \mathbb{E}[e^{-2i\pi\mathbf{f}^T\mathbf{x}}] = \int_{\mathbb{R}} p(\mathbf{x}) e^{-2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{x}$
- Probability distribution: $p(\mathbf{x}) = \int_{\mathbb{R}} \phi_x(\mathbf{f}) e^{+2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{f}$
- Cumulants:
  - Definition: $\ln(\phi_x(\mathbf{f})) = \sum_{n=1}^{+\infty} \frac{(-2i\pi)^n}{n!} \sum_{k_1=1}^{K} \sum_{k_n=1}^{K} \kappa^n_{k_1\dots k_n}[\mathbf{x}] f_{k_1}\dots f_{k_n}$
  - $\kappa^n[\mathbf{x}]$ is an n-th order tensor
  - $\kappa^1[\mathbf{x}]$ is the mean vector, $\kappa^2[\mathbf{x}]$ is the covariance matrix
  - If $p(\mathbf{x})$ is symmetric ($p(-\mathbf{x}) = p(\mathbf{x})$), $\kappa^n[\mathbf{x}] = 0$ for any odd value n
  - the ratio $\kappa^4_{k,k,k,k}[\mathbf{x}]/(\kappa^2_{k,k}[\mathbf{x}])^2$ is called "kurtosis"

9/48 Une école de l'IMT Audio source separation

Real random vectors

- Notation: $\phi[\mathbf{x}]$ denotes a function of $p(\mathbf{x})$
- Mean vector: $\mu_x = \mathbb{E}[\mathbf{x}]$
- Covariance matrix: $\Sigma_{xx} = \mathbb{E}[(\mathbf{x}-\mu_x)(\mathbf{x}-\mu_x)^T]$
- Characteristic function: $\phi_x(\mathbf{f}) = \mathbb{E}[e^{-2i\pi\mathbf{f}^T\mathbf{x}}] = \int_{\mathbb{R}} p(\mathbf{x}) e^{-2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{x}$
- Probability distribution: $p(\mathbf{x}) = \int_{\mathbb{R}} \phi_x(\mathbf{f}) e^{+2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{f}$
- Cumulants:
  - Definition: $\ln(\phi_x(\mathbf{f})) = \sum_{n=1}^{+\infty} \frac{(-2i\pi)^n}{n!} \sum_{k_1=1}^{K} \sum_{k_n=1}^{K} \kappa^n_{k_1\dots k_n}[\mathbf{x}] f_{k_1}\dots f_{k_n}$
  - $\kappa^n[\mathbf{x}]$ is an n-th order tensor
  - $\kappa^1[\mathbf{x}]$ is the mean vector, $\kappa^2[\mathbf{x}]$ is the covariance matrix
  - If $p(\mathbf{x})$ is symmetric ($p(-\mathbf{x}) = p(\mathbf{x})$), $\kappa^n[\mathbf{x}] = 0$ for any odd value n
  - the ratio $\kappa^4_{k,k,k,k}[\mathbf{x}]/(\kappa^2_{k,k}[\mathbf{x}])^2$ is called "kurtosis"

9/48 Une école de l'IMT Audio source separation

Real random vectors

- Notation: $\phi[\mathbf{x}]$ denotes a function of $p(\mathbf{x})$
- Mean vector: $\mu_x = \mathbb{E}[\mathbf{x}]$
- Covariance matrix: $\Sigma_{xx} = \mathbb{E}[(\mathbf{x}-\mu_x)(\mathbf{x}-\mu_x)^T]$
- Characteristic function:
  $\phi_x(\mathbf{f}) = \mathbb{E}[e^{-2i\pi\mathbf{f}^T\mathbf{x}}] = \int_{\mathbb{R}} p(\mathbf{x}) e^{-2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{x}$
- Probability distribution: $p(\mathbf{x}) = \int_{\mathbb{R}} \phi_x(\mathbf{f}) e^{+2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{f}$
- Cumulants:
  - Definition: $\ln(\phi_x(\mathbf{f})) = \sum_{n=1}^{+\infty} \frac{(-2i\pi)^n}{n!} \sum_{k_1=1}^{K} \sum_{k_n=1}^{K} \kappa^n_{k_1\dots k_n}[\mathbf{x}] f_{k_1}\dots f_{k_n}$
  - $\kappa^n[\mathbf{x}]$ is an n-th order tensor
  - $\kappa^1[\mathbf{x}]$ is the mean vector, $\kappa^2[\mathbf{x}]$ is the covariance matrix
  - If $p(\mathbf{x})$ is symmetric ($p(-\mathbf{x}) = p(\mathbf{x})$), $\kappa^n[\mathbf{x}] = 0$ for any odd value n
  - the ratio $\kappa^4_{k,k,k,k}[\mathbf{x}]/(\kappa^2_{k,k}[\mathbf{x}])^2$ is called "kurtosis"

Real random vectors

- Notation: $\phi[\mathbf{x}]$ denotes a function of $p(\mathbf{x})$
- Mean vector: $\mu_x = \mathbb{E}[\mathbf{x}]$
- Covariance matrix: $\Sigma_{xx} = \mathbb{E}[(\mathbf{x}-\mu_x)(\mathbf{x}-\mu_x)^T]$
- Characteristic function:
  $\phi_x(\mathbf{f}) = \mathbb{E}[e^{-2i\pi\mathbf{f}^T\mathbf{x}}] = \int_{\mathbb{R}} p(\mathbf{x}) e^{-2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{x}$
- Probability distribution: $p(\mathbf{x}) = \int_{\mathbb{R}} \phi_x(\mathbf{f}) e^{+2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{f}$
- Cumulants:
  - Definition: $\ln(\phi_x(\mathbf{f})) = \sum_{n=1}^{+\infty} \frac{(-2i\pi)^n}{n!} \sum_{k_1=1}^{K} \sum_{k_n=1}^{K} \kappa^n_{k_1\dots k_n}[\mathbf{x}] f_{k_1}\dots f_{k_n}$
  - $\kappa^n[\mathbf{x}]$ is an n-th order tensor
  - $\kappa^1[\mathbf{x}]$ is the mean vector, $\kappa^2[\mathbf{x}]$ is the covariance matrix
  - If $p(\mathbf{x})$ is symmetric ($p(-\mathbf{x}) = p(\mathbf{x})$), $\kappa^n[\mathbf{x}] = 0$ for any odd value n
  - the ratio $\kappa^4_{k,k,k,k}[\mathbf{x}]/(\kappa^2_{k,k}[\mathbf{x}])^2$ is called "kurtosis"

9/48 Une école de l'IMT Audio source separation

Real random vectors

- Notation: $\phi[\mathbf{x}]$ denotes a function of $p(\mathbf{x})$
- Mean vector: $\mu_x = \mathbb{E}[\mathbf{x}]$
- Covariance matrix: $\Sigma_{xx} = \mathbb{E}[(\mathbf{x}-\mu_x)(\mathbf{x}-\mu_x)^T]$
- Characteristic function:
  $\phi_x(\mathbf{f}) = \mathbb{E}[e^{-2i\pi\mathbf{f}^T\mathbf{x}}] = \int_{\mathbb{R}} p(\mathbf{x}) e^{-2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{x}$
- Probability distribution: $p(\mathbf{x}) = \int_{\mathbb{R}} \phi_x(\mathbf{f}) e^{+2i\pi\mathbf{f}^T\mathbf{x}} d\mathbf{f}$
- Cumulants:
  - Definition: $\ln(\phi_x(\mathbf{f})) = \sum_{n=1}^{+\infty} \frac{(-2i\pi)^n}{n!} \sum_{k_1=1}^{K} \sum_{k_n=1}^{K} \kappa^n_{k_1\dots k_n}[\mathbf{x}] f_{k_1}\dots f_{k_n}$
  - $\kappa^n[\mathbf{x}]$ is an n-th order tensor
  - $\kappa^1[\mathbf{x}]$ is the mean vector, $\kappa^2[\mathbf{x}]$ is the covariance matrix
  - If $p(\mathbf{x})$ is symmetric ($p(-\mathbf{x}) = p(\mathbf{x})$), $\kappa^n[\mathbf{x}] = 0$ for any odd value n
  - the ratio $\kappa^4_{k,k,k,k}[\mathbf{x}]/(\kappa^2_{k,k}[\mathbf{x}])^2$ is called "kurtosis"

9/48 Une école de l'IMT Audio source separation

Real Gaussian random vectors

- The Gaussian distribution is the one such that all cumulants of order $n > 2$ are zero
- Characteristic function
$$\phi_x(\mathbf{f}) = \exp(-2i\pi\mathbf{f}^T\mu_x - 2\pi^2\mathbf{f}^T\Sigma_{xx}\mathbf{f})$$
- Probability density function (defined if Σ_{xx} is invertible)
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{K}{2}}\det(\Sigma_{xx})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_x)^T\Sigma_{xx}^{-1}(\mathbf{x}-\mu_x)\right)$$

Real Gaussian random vectors

- The Gaussian distribution is the one such that all cumulants of order $n > 2$ are zero
- Characteristic function

$$\phi_x(\mathbf{f}) = \exp(-2i\pi\mathbf{f}^T\mu_x - 2\pi^2\mathbf{f}^T\Sigma_{xx}\mathbf{f})$$

- Probability density function (defined if Σ_{xx} is invertible)

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{K}{2}}\det(\Sigma_{xx})^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_x)^T\Sigma_{xx}^{-1}(\mathbf{x}-\mu_x)\right)$$

Real Gaussian random vectors

- The Gaussian distribution is the one such that all cumulants of order $n > 2$ are zero
- Characteristic function

$$\phi_x(\mathbf{f}) = \exp(-2i\pi\mathbf{f}^T\mu_x - 2\pi^2\mathbf{f}^T\Sigma_{xx}\mathbf{f})$$

- Probability density function (defined if Σ_{xx} is invertible)

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{K}{2}}\det(\Sigma_{xx})^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_x)^T\Sigma_{xx}^{-1}(\mathbf{x}-\mu_x)\right)$$

10/48 Une école de l'IMT Audio source separation

WSS vector processes

- Definition: the cumulants of orders 1 et 2 are translation-invariant
- Covariance matrices of 2 centered WSS processes $\mathbf{x}(t)$ and $\mathbf{y}(t)$:
  - Definition: $\mathbf{R}_{xy}(\tau) = \mathbb{E}\left[\mathbf{x}(t+\tau)\mathbf{y}(t)^T\right]$
  - Property: $\mathbf{R}_{xx}(0) = \Sigma_{xx}$ is Hermitian and positive semi-definite.
- PSD matrices of a WSS process $\mathbf{x}(t)$:
  - Definition: $\mathbf{S}_{xx}(\nu) = \sum_{\tau\in\mathbb{Z}} \mathbf{R}_{xx}(\tau)e^{-2i\pi\nu\tau}$
  - Property: $\forall\nu$, $\mathbf{S}_{xx}(\nu)$ is Hermitian and positive semi-definite

WSS vector processes

- Definition: the cumulants of orders 1 et 2 are translation-invariant
- Covariance matrices of 2 centered WSS processes $\mathbf{x}(t)$ and $\mathbf{y}(t)$:
  - Definition: $\mathbf{R}_{xy}(\tau) = \mathbb{E}\left[\mathbf{x}(t+\tau)\mathbf{y}(t)^T\right]$
  - Property: $\mathbf{R}_{xx}(0) = \Sigma_{xx}$ is Hermitian and positive semi-definite.
- PSD matrices of a WSS process $\mathbf{x}(t)$:
  - Definition: $\mathbf{S}_{xx}(\nu) = \sum_{\tau\in\mathbb{Z}} \mathbf{R}_{xx}(\tau)e^{-2i\pi\nu\tau}$
  - Property: $\forall\nu$, $\mathbf{S}_{xx}(\nu)$ is Hermitian and positive semi-definite

11/48 Une école de l'IMT Audio source separation

WSS vector processes

- Definition: the cumulants of orders 1 et 2 are translation-invariant
- Covariance matrices of 2 centered WSS processes $\mathbf{x}(t)$ and $\mathbf{y}(t)$:
  - Definition: $\mathbf{R}_{xy}(\tau) = \mathbb{E}\left[\mathbf{x}(t+\tau)\mathbf{y}(t)^T\right]$
  - Property: $\mathbf{R}_{xx}(0) = \Sigma_{xx}$ is Hermitian and positive semi-definite.
- PSD matrices of a WSS process $\mathbf{x}(t)$:
  - Definition: $\mathbf{S}_{xx}(\nu) = \sum_{\tau \in \mathbb{Z}} \mathbf{R}_{xx}(\tau) e^{-2i\pi\nu\tau}$
  - Property: $\forall \nu$, $\mathbf{S}_{xx}(\nu)$ is Hermitian and positive semi-definite

Information theory

- Shannon entropy
  - Definition: $\mathbb{H}[\mathbf{x}] = -\mathbb{E}[\ln(p(\mathbf{x}))]$
  - $\mathbb{H}[\mathbf{x}]$ is not necessarily non-negative for a continuous r.v.
- Kullback-Leibler divergence
  - $D_{KL}(p||q) = \int p(\mathbf{x}) \ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x}$
  - Property: $D_{KL}(p||q) \geq 0$, $D_{KL}(p||q) = 0$ if and only if $p = q$
- Mutual information
  - Definition: $\mathbb{I}[\mathbf{x}] = \mathbb{E}\left[\ln\left(\frac{p(\mathbf{x})}{p(x_1)\dots p(x_K)}\right)\right] = D_{KL}(p(\mathbf{x})||p(x_1)\dots p(x_K))$
  - Property: $\mathbb{I}[\mathbf{x}] = 0$ if and only if $x_1 \dots x_K$ are mutually independent
  - Relationship with entropy: $\mathbb{I}[\mathbf{x}] = \sum_{k=1}^K \mathbb{H}[x_k] - \mathbb{H}[\mathbf{x}]$

Information theory

- Shannon entropy
  - Definition: $\mathbb{H}[\mathbf{x}] = -\mathbb{E}[\ln(p(\mathbf{x}))]$
  - $\mathbb{H}[\mathbf{x}]$ is not necessarily non-negative for a continuous r.v.
- Kullback-Leibler divergence
  - $D_{KL}(p||q) = \int p(\mathbf{x}) \ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x}$
  - Property: $D_{KL}(p||q) \geq 0$, $D_{KL}(p||q) = 0$ if and only if $p = q$
- Mutual information
  - Definition: $\mathbb{I}[\mathbf{x}] = \mathbb{E}\left[\ln\left(\frac{p(\mathbf{x})}{p(x_1)\dots p(x_K)}\right)\right] = D_{KL}(p(\mathbf{x})||p(x_1)\dots p(x_K))$
  - Property: $\mathbb{I}[\mathbf{x}] = 0$ if and only if $x_1 \dots x_K$ are mutually independent
  - Relationship with entropy: $\mathbb{I}[\mathbf{x}] = \sum_{k=1}^K \mathbb{H}[x_k] - \mathbb{H}[\mathbf{x}]$

Information theory

- Shannon entropy
  - Definition: $\mathbb{H}[\mathbf{x}] = -\mathbb{E}[\ln(p(\mathbf{x}))]$
  - $\mathbb{H}[\mathbf{x}]$ is not necessarily non-negative for a continuous r.v.
- Kullback-Leibler divergence
  - $D_{KL}(p||q) = \int p(\mathbf{x}) \ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x}$
  - Property: $D_{KL}(p||q) \geq 0$, $D_{KL}(p||q) = 0$ if and only if $p = q$
- Mutual information
  - Definition: $\mathbb{I}[\mathbf{x}] = \mathbb{E}\left[\ln\left(\frac{p(\mathbf{x})}{p(x_1)\dots p(x_K)}\right)\right] = D_{KL}(p(\mathbf{x})||p(x_1)\dots p(x_K))$
  - Property: $\mathbb{I}[\mathbf{x}] = 0$ if and only if $x_1 \dots x_K$ are mutually independent
  - Relationship with entropy: $\mathbb{I}[\mathbf{x}] = \sum_{k=1}^K \mathbb{H}[x_k] - \mathbb{H}[\mathbf{x}]$

Part III

Linear instantaneous mixtures

Blind source separation (BSS)

- Observation model:
  - $\forall t,\ \mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ where $\mathbf{A} \in \mathbb{R}^{M\times K}$ is called the "mixture matrix"
  - Sources are assumed IID: $p(\{s_k(t)\}_{k,t}) = \prod_{k=1}^{K}\prod_{t=1}^{T} p_k(s_k(t))$
- Problem: estimate $\mathbf{A}$ and sources $\mathbf{s}(t)$ given $\mathbf{x}(t)$
- Definition: non-mixing matrix
  - a matrix $\mathbf{C}$ of dimension $K \times K$ is non-mixing if and only if it has a unique non-zero entry in each row and each column
- If $\widetilde{\mathbf{s}}(t) = \mathbf{C}\mathbf{s}(t)$ and $\widetilde{\mathbf{A}} = \mathbf{A}\,\mathbf{C}^{-1}$, then $\mathbf{x}(t) = \widetilde{\mathbf{A}}\widetilde{\mathbf{s}}(t)$ is another admissible decomposition of the observations
  - Sources can be recovered up to a permutation and a multiplicative factor

Linear separation of sources

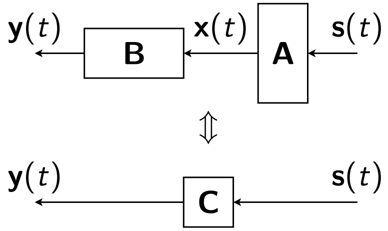
- Let $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t)$, where $\mathbf{B} \in \mathbb{R}^{K\times M}$ is referred to as the "separation matrix"
- Linear separation is feasible if $\mathbf{A}$ has rank K:
  - We get $\mathbf{y}(t) = \mathbf{s}(t)$ by defining:
    - $\mathbf{B} = \mathbf{A}^{-1}$ in the determined case ($M = K$)
    - $\mathbf{B} = \mathbf{A}^{\dagger}$ in the over-determined case ($M > K$)
  - the pseudo-inverse $\mathbf{A}^{\dagger} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is such that $\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{I}_K$
- In the under-determined case ($M < K$), separation is not feasible

Part IV

Independent component analysis

Independent component analysis (ICA)

- In practice matrix $\mathbf{A}$ is unknown:
  - We look for a matrix $\mathbf{B}$ that makes the y_k independent (ICA)
  - We then get equation $\mathbf{y}(t) = \mathbf{Cs}(t)$, where $\mathbf{C} = \mathbf{BA}$
  - The problem is solved if matrix $\mathbf{C}$ is non-mixing



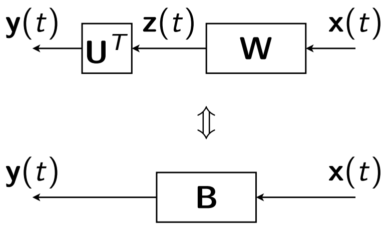
- Theorem (identifiability)
  - Let s_k be K IID sources, among which at most one is Gaussian, and $\mathbf{y}(t) = \mathbf{Cs}(t)$ with $\mathbf{C}$ invertible ((over)-determined case). If signals $y_k(t)$ are independent, then $\mathbf{C}$ is non-mixing.

Whitening

- We now suppose that the sources are centered: $\mathbb{E}[\mathbf{s}(t)] = \mathbf{0}$ and that the mixture is (over-)determined
- Canonical problem: we can assume without loss of generality that $\mathbf{s}(t)$ is spatially white ($\Sigma_{ss} = \mathbb{E}[\mathbf{s}(t)\mathbf{s}(t)^T] = \mathbf{I}_K$)
- Then $\Sigma_{xx} = \mathbf{A}\Sigma_{ss}\mathbf{A}^T = \mathbf{A}\,\mathbf{A}^T$: $\mathbf{A}$ is a matrix square root of Σ_{xx}
- We first aim to whiten (decorrelate) the mixture:
  - Σ_{xx} is diagonalizable in an orthonormal basis: $\Sigma_{xx} = \mathbf{Q}\Lambda^2\mathbf{Q}^T$ where $\Lambda = \mathrm{diag}(\lambda_1 \dots \lambda_M)$ with $\lambda_1 \geq \lambda_K > \lambda_{K+1} = \lambda_M = 0$ (the rank of Σ_{xx} is equal to K)
  - Let $\mathbf{S} = \mathbf{Q}_{(:,1:K)}\Lambda_{(1:K,1:K)} \in \mathbb{R}^{M\times K}$
  - $\mathbf{S}$ is a matrix square root of Σ_{xx}: $\Sigma_{xx} = \mathbf{S}\,\mathbf{S}^T$
  - Let $\mathbf{W} = \mathbf{S}^\dagger$ and $\mathbf{z}(t) = \mathbf{Wx}(t)$
  - Then $\mathbf{z}(t)$ is white ($\mathbb{E}[\mathbf{z}(t)] = \mathbf{0}$ and $\Sigma_{zz} = \mathbf{W}\Sigma_{xx}\mathbf{W}^T = \mathbf{I}$)

Whitening

- We conclude without loss of generality that $\mathbf{U} \triangleq \mathbf{WA}$ is a rotation matrix ($\mathbf{UU}^T = \mathbf{I}$).
- Then $\mathbf{y}(t) = \mathbf{U}^T\mathbf{z}(t) = \mathbf{U}^T\mathbf{Wx}(t) = (\mathbf{WA})^{-1}(\mathbf{WA})\mathbf{s}(t) = \mathbf{s}(t)$.
- We can thus assume $\mathbf{B} = \mathbf{U}^T\mathbf{W}$ where $\mathbf{U}$ is a rotation matrix.



Higher order statistics

- One can estimate Σ_{xx} from the observations and get $\mathbf{W}$
- The whiteness property (second order cumulants) determines $\mathbf{W}$ and leaves $\mathbf{U}$ unknown.
- If sources are Gaussian, the z_k are independent and $\mathbf{U}$ cannot be determined.
- In order to determine rotation $\mathbf{U}$, we need to exploit the non-Gaussianity of sources and characterize the independence property by using cumulants of order greater than 2.

Contrast functions

- Definition: ϕ is a "contrast function" if and only if $\phi[\mathbf{Cs}(t)] \geq \phi[\mathbf{s}(t)]$ $\forall \mathbf{C}$ and if $\phi[\mathbf{Cs}(t)] = \phi[\mathbf{s}(t)] \Leftrightarrow \mathbf{C}$ is non-mixing.
- Separation is performed by minimizing $\phi[\mathbf{y}(t) = \mathbf{Cs}(t)]$ with respect to $\mathbf{U}$ (or $\mathbf{B}$)
- "Canonical" contrast function: $\phi_{IM}[\mathbf{y}(t)] = \mathbb{I}[\mathbf{y}(t)]$
- Orthogonal contrasts: to be minimized under the constraint $\mathbb{E}[\mathbf{y}(t)\mathbf{y}(t)^T] = \mathbf{I}$. For instance, $\phi^{\circ}_{IM}[\mathbf{y}(t)] = \sum_{k=1}^{K} \mathbb{H}(y_k(t))$
- Order 4 approximation of ϕ°_{IM}: $\phi^{\circ}_{ICA}[\mathbf{y}(t)] = \sum_{ijkl \neq iiii} (\kappa^4_{ijkl}[\mathbf{y}(t)])^2$
- Descent algorithms for minimizing ϕ with respect to $\mathbf{B}$ or $\mathbf{U}$:
  - Gradient algorithm applied to matrix $\mathbf{B}$
  - Parameterization of $\mathbf{U}$ with Givens rotations and coordinate descent

Algorithm

1. Estimation of the covariance matrix Σ_{xx}
2. Diagonalization of Σ_{xx}: $\Sigma_{xx} = \mathbf{Q}\Lambda^2\mathbf{Q}^T$ where $\Lambda = \mathrm{diag}(\lambda_1 \ldots \lambda_M)$ with $\lambda_1 \geq \ldots \geq \lambda_M \geq 0$
3. Computation of $\mathbf{S} = \mathbf{Q}_{(:,1:K)}\Lambda_{(1:K,1:K)}$
4. Computation of the whitening matrix $\mathbf{W} = \mathbf{S}^{\dagger}$
5. Data whitening: $\mathbf{z}(t) = \mathbf{Wx}(t)$
6. Estimation of $\mathbf{U}$ by minimizing the contrast function ϕ°
7. Estimation of source signals via $\mathbf{y}(t) = \mathbf{U}^T\mathbf{z}(t)$

Part V

Second order methods

Temporal coherence of sources

- Model: $\mathbb{E}(\mathbf{s}(t)) = \mathbf{0}$, $\mathbf{R}_{ss}(\tau) = \mathbb{E}\left(\mathbf{s}(t+\tau)\mathbf{s}(t)^T\right) = \mathrm{diag}(r_{s_k}(\tau))$
- Canonical problem: we assume that $\Sigma_{ss} = \mathbf{R}_{ss}(0) = \mathbf{I}$
- We first aim to spatially whiten the mixture:
  - Let $\mathbf{S}$ be a matrix square root of Σ_{xx}
  - Let $\mathbf{W} = \mathbf{S}^{\dagger}$ and $\mathbf{z}(t) = \mathbf{Wx}(t)$
- Since $\Sigma_{xx} = \mathbf{A}\mathbf{A}^T$, $\mathbf{U} \triangleq \mathbf{WA}$ is a rotation matrix
- However, $\forall \tau \in \mathbb{Z}$, $\mathbf{R}_{zz}(\tau) = \mathbf{UR}_{ss}(\tau)\mathbf{U}^T$
- The joint diagonalization of matrices $\mathbf{R}_{zz}(\tau)$ for various values of τ permits us to identify rotation $\mathbf{U}$

Joint diagonalization

- Unicity theorem :
  - Let a set of matrices $\mathbf{R}_{zz}(\tau)$ of dimension $K \times K$ and of the form $\mathbf{R}_{zz}(\tau) = \mathbf{U}\mathbf{R}_{ss}(\tau)\mathbf{U}^T$ with $\mathbf{U}$ unitary and $\mathbf{R}_{ss}(\tau) = \mathrm{diag}(r_{s_k}(\tau))$. Then $\mathbf{U}$ is unique (up to a non-mixing matrix) if and only if $\forall 1 \leq k \neq l \leq K$, there is τ such that $r_{s_k}(\tau) \neq r_{s_l}(\tau)$
- Joint diagonalization methods: minimize the criterion

$$J(\mathbf{U}) = \sum_\tau \|\mathbf{U}^T\mathbf{R}_{zz}(\tau)\mathbf{U} - \mathrm{diag}(\mathbf{U}^T\mathbf{R}_{zz}(\tau)\mathbf{U})\|_F^2$$

  - Parameterization of $\mathbf{U}$ with Givens rotations and coordinate descent

SOBI algorithm

- *Second Order Blind Identification (SOBI)*
  1. Estimation and diagonalization of Σ_{xx}: $\Sigma_{xx} = \mathbf{Q}\Lambda^2\mathbf{Q}^T$ where $\Lambda = \mathrm{diag}(\lambda_1 \ldots \lambda_M)$ with $\lambda_1 \geq \ldots \geq \lambda_M \geq 0$
  2. Computation of $\mathbf{S} = \mathbf{Q}_{(:,1:K)}\Lambda_{(1:K,1:K)}$
  3. Computation of the whitening matrix $\mathbf{W} = \mathbf{S}^\dagger$
  4. Data whitening: $\mathbf{z}(t) = \mathbf{W}\mathbf{x}(t)$
  5. Estimation of covariance matrices $\mathbf{R}_{zz}(\tau)$ for various delays τ
  6. Approximate joint diagonalization of matrices $\mathbf{R}_{zz}(\tau)$ in a common basis $\mathbf{U}$
  7. Estimation of source signals via $\mathbf{y}(t) = \mathbf{U}^T\mathbf{z}(t)$

Non-stationarity of sources

- Model: $\mathbb{E}(\mathbf{s}(t)) = \mathbf{0}$, $\Sigma_{ss}(t) \triangleq \mathbb{E}\left(\mathbf{s}(t)\mathbf{s}(t)^T\right) = \mathrm{diag}(\sigma_k^2(t))$
- Then $\forall t \in \mathbb{Z}$, $\Sigma_{xx}(t) = \mathbf{A}\Sigma_{ss}(t)\mathbf{A}^T$
- Joint diagonalization methods: minimize the criterion

$$J(\mathbf{B}) = \sum_t \|\mathbf{B}\Sigma_{xx}(t)\mathbf{B}^T - \mathrm{diag}(\mathbf{B}\Sigma_{xx}(t)\mathbf{B}^T)\|_F^2$$

  - Gradient descent algorithm applied to matrix $\mathbf{B}$
  - In the over-determined case, $\mathbf{B}$ must be constrained to span the principal subspace of all matrices $\Sigma_{xx}(t)$
- Variant of the SOBI algorithm:
  1. Segmentation of source signals and estimation of covariance matrices $\Sigma_{xx}(t)$ on windows centered at different times t
  2. Joint diagonalization of matrices $\Sigma_{xx}(t)$ in a common basis $\mathbf{B}$
  3. Estimation of source signals via $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t)$

Conclusion of the first part

- The use of higher order cumulants is only necessary for the non-Gaussian IID source model
- Second order statistics are sufficient for sources that are:
  - stationary but not IID ($\rightarrow$ spectral dynamics)
  - non stationary ($\rightarrow$ temporal dynamics)
- Remember that classical tools (based on second order statistics) are appropriate for blind separation of independent (and possibly Gaussian) sources, on condition that the spectral / temporal source dynamics is taken into account.