A Diagonal Plus Low-Rank Covariance Model for Computationally Efficient Source Separation

Antoine Liutkus (INRIA, France)
Kazuyoshi Yoshii (Kyoto University/RIKEN AIP, Japan)
Outline

• We introduce positive semidefinite tensor factorization (PSDTF) based on the Log-Det divergence
  ▪ A natural extension of nonnegative matrix factorization (NMF) based on the Itakura-Saito divergence
  ▪ Estimation of locally-stationary Gaussian processes

• We propose a constrained version of LD-PSDTF for reducing computational complexity
  ▪ Kernel matrices are restricted to diagonal + low-rank matrices
  ▪ Woodbury formula is used for inversing kernel matrices
**Background**

- **Source separation is essential for various applications**
  - Speech recognition and understanding
  - Automatic music transcription
- **Phase information has not been used in most studies**
  - The characteristics of sounds can be represented well in the power domain by discarding the phase information
  - The low-rankness and sparseness are useful clues

---

**Nonnegative matrix factorization (NMF)**

- Interpretable
- Easy to implement
- Efficient

---

**Diagram**

- **Signal**
  - **FFT**
  - **Inverse FFT**
  - Taking energy
  - **Real**
  - **Imag**
  - **Complex spectrum**
  - **Power spectrum**

- **Power**
  - **FFT**
  - **Inverse FFT**
  - Taking energy
  - **Real**
  - **Imag**
  - **Complex spectrum**
  - **Power spectrum**

---
Motivation

- Phase-aware source separation is promising
  - NMF can be extended based on **additivity of complex spectra**

<table>
<thead>
<tr>
<th></th>
<th>Frequency bins</th>
<th>Time frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex NMF [Kameoka 2009]</td>
<td>Independent</td>
<td>Independent</td>
</tr>
<tr>
<td>High Resolution NMF [Badeau 2011]</td>
<td>Independent</td>
<td>Autoregressive</td>
</tr>
<tr>
<td>PSDTF [Yoshii 2013]</td>
<td>Correlated</td>
<td>Independent</td>
</tr>
</tbody>
</table>

Complex value

\[ x_{ft} = r_{ft} (\cos \theta_{ft} + i \sin \theta_{ft}) \]

The values of magnitude and phase are not determined independently at frequency bins.
• Each observed nonnegative vector is approximated as the weighted sum of basis nonnegative vectors

\[ \mathbf{x}_f \mathbf{f} = \text{diag}(\mathbf{w}_f \mathbf{H}) \]

\[ \mathbf{y}_f = \sum_{k=1}^{K} h_{k} \mathbf{w}_k \]

Minimize \( D_{IS}(\mathbf{x}_t | \mathbf{y}_t) \)

Calculate the power spectrum at each frame

\( \hat{x}_t = \text{diag}(\mathbf{x}_t \mathbf{x}_t^H) \)
Log-Det PSDTF (LD-PSDTF) [Yoshii 2013]

- Each observed pos. semidef. matrix is approximated as the weighted sum of basis pos. semidef. matrices

$$\mathbf{X}_f = \mathbf{x}_f \mathbf{x}_f^H$$

$$\mathbf{Y}_t = \sum_{k=1}^{K} h_{kt} \mathbf{V}_k$$

Calculate the covariance matrix at each frame

PSD matrices

Basis matrices

Activations

Nonnegative vectors

Complex spectrogram

Minimize $$D_{LD}(\mathbf{X}_t|\mathbf{Y}_t)$$
IS-NMF vs LD-PSDTF

- LD-PSDTF is a natural extension of IS-NMF
  - Nonnegativity of scalars $\rightarrow$ Positive semidefiniteness of matrices
  - PSDTF reduces to NMF when all PSD matrices are diagonal

\[ \hat{x}_t \approx \sum_{k=1}^{K} h_{kt} w_k \]

\[ X_t \approx \sum_{k=1}^{K} h_{kt} V_k \]
**Itakura-Saito NMF (IS-NMF)** [Févotte 2009]

- NMF based on the Itakura-Saito divergence
  - The mixture spectrogram is approximated as a low-rank matrix
  - The number of sources $K$ should be specified in advance

$$\hat{x}_{ft} \approx \hat{y}_{ft} = \sum_{k=1}^{K} w_{kf} h_{kt}$$

$$D_{IS}(\hat{x}_{ft} | \hat{y}_{ft}) = -\log \frac{\hat{x}_{ft}}{\hat{y}_{ft}} + \frac{\hat{x}_{ft}}{\hat{y}_{ft}} - 1$$

Scale-invariant measure

$$D_{IS}(\alpha \hat{x}_{ft} | \alpha \hat{y}_{ft}) = D_{IS}(\hat{x}_{ft} | \hat{y}_{ft})$$
Log-Det PSDTF (LD-PSDTF) [Yoshii 2013]

- PSDTF based on the log-determinant divergence
  - The covariance matrix at each frame is approximated as the weighted sum of covariance matrices (basis matrices)

\[
X_t \approx Y_t = \sum_{k=1}^{K} V_k h_{kt}
\]

\[
D_{LD}(X_t | Y_t) = -\log |X_t Y_t^{-1}| + \text{tr}(X_t Y_t^{-1}) - F
\]

Observed spectrogram

Reconstructed spectrogram

Scale-invariant measure

\[
D_{LD}(X_t | Y_t) = D_{LD}(\alpha X_t | \alpha Y_t)
\]
Probabilistic Formulation

• The source signals are assumed to follow independent locally-stationary Gaussian processes in the time domain

  ▪ A mixture signal is the sum of multiple source signals

Assume the signals to be stationary in a short window

\[
x_t = x_{1t} + x_{2t} + x_{3t}
\]

\[
x_{1t} \sim N_c(0, h_{1t}V_1)
\]

\[
x_{2t} \sim N_c(0, h_{2t}V_2)
\]

\[
x_{3t} \sim N_c(0, h_{3t}V_3)
\]
Mixing Process & Demixing Process

- **Sum of Gaussian variables → Gaussian variable**
  \[ x_{1t} \sim N_c(0, Y_{1t}) \]
  \[ x_{2t} \sim N_c(0, Y_{2t}) \]
  \[ x_{3t} \sim N_c(0, Y_{3t}) \]
  \[ x_t = x_{1t} + x_{2t} + x_{3t} \]
  \[ \sim N_c(0, Y_{1t} + Y_{3t} + Y_{3t} = Y_t) \]

- **Gaussian variable → Sum of Gaussian variables**
  \[ x_{1t} \sim N_c(0, Y_{1t}) \]
  \[ x_{2t} \sim N_c(0, Y_{2t}) \]
  \[ x_{3t} \sim N_c(0, Y_{3t}) \]
  \[ x_t = x_{1t} + x_{2t} + x_{3t} \]
  \[ \sim N_c(0, Y_{1t} + Y_{3t} + Y_{3t} = Y_t) \]

\[ x_{1t}|x_t \sim N_c(Y_{1t}Y_t^{-1}x_t, Y_{1t} - Y_{1t}Y_t^{-1}Y_{1t}) \]
\[ x_{2t}|x_t \sim N_c(Y_{2t}Y_t^{-1}x_t, Y_{2t} - Y_{2t}Y_t^{-1}Y_{2t}) \]
\[ x_{3t}|x_t \sim N_c(Y_{3t}Y_t^{-1}x_t, Y_{3t} - Y_{3t}Y_t^{-1}Y_{3t}) \]

All the frequency components of each source spectrum can be estimated jointly via Wiener filtering.
Maximum Likelihood Estimation

- We aim to estimate $H, V$ that maximizes the likelihood

$$ x_t \sim N_c \left( 0, \sum_{k=1}^{K} h_{kt} V_k \right) \rightarrow \text{Maximize} $$

**Observed complex spectrogram**

$$ X_t = x_t x_t^H $$

**Observed covariance matrix**

$$ Y_t = \sum_{k=1}^{K} h_{kt} V_k $$

**Approx. covariance matrix**

**Gaussian log-likelihood**

$$ \log p(X_t|Y_t) = -\frac{1}{2} \log|Y_t| - \frac{1}{2} \text{tr}(X_t Y_t^{-1}) \rightarrow \text{Maximize} $$

**Log-Det divergence**

$$ D(X_t|Y_t) = -\log|X_t Y_t^{-1}| + \text{tr}(X_t Y_t^{-1}) - F \rightarrow \text{Minimize} $$

Equivalent!
Generalized EM Algorithm (Proposed)

- Iteratively update latent sources and parameters
  - Expectation step
    - Calculate covariance matrices
      \[ Y_{kt} = h_{kt} V_k \]
    - Calculate posteriors of source spectra
      \[ x_{kt} | x_t \sim N_c \left( Y_{kt} Y_t^{-1} x_t, Y_{kt} - Y_{kt} Y_t^{-1} Y_{kt} \right) \]
    - Calculate second-order statistics
      \[ E[x_{kt} x_{kt}^H] = E[x_{kt}] E[x_{kt}^H] + V[x_{kt}] \]
  - Maximization step
    - Update parameters (depend on each other)
      \[ h_{kt} \leftarrow \frac{\text{tr} \left( V_k^{-1} E[x_{kt} x_{kt}^H] \right)}{F} \]
      \[ V_k \leftarrow \frac{\sum_{t=1}^{T} h_{kt}^{-1} E[x_{kt} x_{kt}^H]}{T} \]

IS-NMF: \( O(KTF) \)
LD-PSDTF: \( O(KTF^3) \)
Computational Bottleneck

- Inversion of big matrices is computationally prohibitive
  - E step: updating source spectra
    \[ E \left[ x_{kt} x_{kt}^H \right] = Y_{kt} Y_t^{-1} x_t + Y_{kt} - Y_{kt} Y_t^{-1} Y_{kt} \]
  - M step: updating parameters
    \[ h_{kt} \leftarrow \frac{\text{tr} \left( V_k^{-1} E \left[ x_{kt} x_{kt}^H \right] \right)}{F} \]
    \[ V_k \leftarrow \frac{\sum_{t=1}^{T} h_{kt}^{-1} E \left[ x_{kt} x_{kt}^H \right]}{T} \]

The inverse matrices \( Y_t^{-1} \) and \( V_k^{-1} \in \mathbb{C}^{F \times F} \) are required: \( O(F^3) \)

How to calculate these inversions in a more efficient manner?
Covariance Matrix Revisited

• Basis covariance matrices have diagonal + grid patterns
  ▪ Especially for complex spectra with harmonic structures

\[ F \]

\[ V_k \]

Strong correlations between harmonic partials in a lower frequency range

Power spectral densities of harmonic partials (basis vector of NMF)
Covariance Approximation (Proposed)

- Each $V_k$ is approximated as a diagonal + low-rank matrix
  - The rank $N$ can be around the number of harmonic partials

$$V_k = [w_k] + L_k [s_k] L_k^H$$

- Basis covariance matrix
- Basis power spectrum
- $V_k \in C^{F \times F}$
- $s_k \in C^N$
- $w_k \in C^F$
- $L_k \in C^{F \times N}$
EM Algorithm Revisited

- The inversion of big matrices are required
  - E step: updating source spectra
    \[ E \left[ x_{kt} x_{kt}^H \right] = Y_{kt} Y_t^{-1} x_t + Y_{kt} - Y_{kt} Y_t^{-1} Y_{kt} \]
  - M step: updating parameters
    \[ h_{kt} \leftarrow \frac{\text{tr} \left( V_k^{-1} E \left[ x_{kt} x_{kt}^H \right] \right)}{F} \]
    \[ V_k \leftarrow \frac{\sum_{t=1}^{T} h_{kt}^{-1} E \left[ x_{kt} x_{kt}^H \right]}{T} \]
    \[ V_k = [w_k] + L_k [s_k] L_k^H \]
    \[ Y_t = \sum_{k=1}^{K} h_{kt} V_k = \sum_{k=1}^{K} h_{kt} [w_k] + \sum_{k=1}^{K} h_{kt} L_k [s_k] L_k^H \]

Each term can be inverted efficiently
Efficient Matrix Inversion

• Use Woodbury formula for covariance matrices

\[
(A + UCV)^{-1} = A^{-1} - A^{-1} U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}
\]

This formula is useful when \(A\) and \(C\) can be inverted efficiently

\[
V_k = [w_k] + L_k [s_k] L_k^H \quad \text{Diagonal matrices!}
\]

\[
V_k^{-1} = [w_k]^{-1} - [w_k]^{-1} L_k \left( [s_k]^{-1} + L_k^H [w_k]^{-1} L_k \right)^{-1} L_k^H [w_k]^{-1} \quad \text{Inversion of a compact matrix!}
\]
Recursive Matrix Inversion

- Use Woodbury formula in a recursive manner

\[
Y_t = \sum_{k=1}^{K} h_{kt}[w_k] + \sum_{k=1}^{K} h_{kt}L_k[s_k]L_k^H \quad \Rightarrow \quad Y_t^{-1}
\]

\[
Y_t^{(p)} \overset{\text{def}}{=} \sum_{k=1}^{K} h_{kt}[w_k] + \sum_{k=1}^{p} h_{kt}L_k[s_k]L_k^H = Y_t^{(p-1)} + h_{pt}L_p[s_p]L_p^H
\]

\[
\left(Y_t^{(p)}\right)^{-1} = \left(Y_t^{(p-1)}\right)^{-1} \quad N \times N
\]

\[- \left(Y_t^{(p-1)}\right)^{-1} L_p \left( h_{pt}^{-1}[s_p]^{-1} + L_p^H \left( Y_t^{(p-1)}\right)^{-1} L_p \right)^{-1} L_p^H \left( Y_t^{(p-1)}\right)^{-1} \]

Recurrence formula starting at \( Y_t^{(0)} = \sum_{k=1}^{K} h_{kt}[w_k] \) (NMF)
Evaluation

• Separation performance vs covariance approximation
  ▪ Synthesize a mixture signal sampled at 16 [kHz]
    ▪ $K = 3$ (C4, E4, G4, piano) • $F = 256$, $T = 840$
  ▪ Test “fast” PSDTF with $N = 0$ (NMF), 1, 5, 10, 50, 256 (PSDTF)
  ▪ Use BSS Eval Toolbox [Vincent2006]

![Waveform](image1)

<table>
<thead>
<tr>
<th>Mixture signal</th>
<th>C</th>
<th>E</th>
<th>G</th>
<th>C+E</th>
<th>C+G</th>
<th>E+G</th>
<th>C+E+G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source signal 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source signal 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source signal 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Separation Performance

- PSDTF ($N = 50$) was comparable with PSDTF ($N = 256$)
Estimated Results

- The off-diagonal elements of each $V_k$ (inter-frequency correlations) can be approximated by a low-rank matrix
  - A limited number of eigenvalues are significantly larger than 0

\[ L_k \]
\[ [s_k] \]
\[ N = 5 \]
\[ N = 10 \]
\[ N = 50 \]
Conclusion

• We introduced positive semidefinite tensor factorization (PSDTF) based on the Log-Det divergence
  ▪ A natural extension of nonnegative matrix factorization (NMF) based on the Itakura-Saito divergence
  ▪ Estimation of locally-stationary Gaussian processes

• We proposed a constrained version of LD-PSDTF for reducing computational complexity
  ▪ Kernel matrices are restricted to diagonal + low-rank matrices
  ▪ Woodbury formula is used for inversing kernel matrices (in a recursive manner)

\[
V_k = [w_k] + L_k [s_k] L_k^H = +
\]