An Inverse-Gamma Source Variance Prior With Factorized Parametrization for Audio Source Separation

Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot, Radu Horaud
Problem: \( J \) Source signals, mixed with filters and summed, are recorded at \( I \) microphones: Recover the original sources!

An ill-posed problem: very large number of unknown variables and parameters.
Problem Formulation in STFT domain

- Separate a mixture of $J$ sources with $I$ microphones.
- In STFT domain the problem becomes:

$$x_{f\ell} = A_{fsf\ell} + b_{f\ell}$$

- $x_{f\ell}$: mixture $[I \times 1]$ observed
- $A_{fsf\ell}$: mixing matrix $[I \times J]$, unknown!
- $b_{f\ell}$: sensor noise $[I \times 1]$, unknown!
- $s_{f\ell}$: source STFT $[J \times 1]$, unknown!

• There are multitudinous MASS methods.
• We embrace the family of methods based on Wiener demixing.
• The general recipe is:
  • Estimate $|s_{j,f,\ell}|^2$, e.g. via NMF\textsuperscript{[1]}.
  • Estimate the mixing matrices $A_f$.
  • Construct demixing Wiener Filters to extract $s_{f,\ell}$ from $x_{f,\ell}$.
  • Iterate ..

\[\text{[Ozerov and Févotte, 2010]}\]
Local Gaussian Composite Model

• Inspired by[1][2]:

• Each source $s_{j,f\ell}$: sum of latent components

$$s_{j,f\ell} = \sum_{k=1}^{K_j} c_{k,f\ell} \iff s_{f\ell} = Gc_{f\ell},$$

with a known binary matrix $G \in \mathbb{N}^{J \times K}$;

• in total we have $K = \sum_{j=1}^{J} K_j$ components.

• Each component follows $p(c_{k,f\ell}) = \mathcal{N}_c(c_{k,f\ell}; 0, u_{k,f\ell})$.

[1][A. Ozerov and C. Févotte, 2010]
Non-Negative Matrix Factorisation (NMF)

- Typically: $u_{k,f\ell} = w_{fk} h_{k\ell}$ as in\cite{Ozerov2010,Arberet2010}
- This is equivalent with NMF on $|s_{j,f\ell}|^2$:

**Benefits:**
- Reduces the number of parameters to be estimated.
- Avoids the permutation of sources between frequencies.

**Limitations:**
- $u_{k,f\ell}$ is of rank=1 (thus $|s_{j,f\ell}|^2$ is of rank=\(|\mathcal{K}_j|\));
- Limited flexibility of the estimated demixing Wiener-filters due to low-rank constraint on $|s_{j,f\ell}|^2$.

---

\cite{Ozerov2010}[A. Ozerov and C. Févotte, 2010]
\cite{Arberet2010}[S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vanderghynst, 2010]
Our Goal

- We would like to have $|s_{j,f\ell}|^2$ be full-rank (i.e. unfactorised);
- use no more parameters as the standard NMF;
- and without introducing frequency-permuation;
- **We want NMF but without factorisation! How?**
Proposed Model Formulation

- Each $u_{k,\ell} \in \mathbb{R}_+$ is considered as a r.v.

$$p(u_{k,\ell}) = IG(\gamma_k, \delta_{k,\ell})$$

$$= \frac{(\delta_{k,\ell})^{\gamma_k}}{\Gamma(\gamma_k)} u_{k,\ell}^{-(\gamma_k+1)} \exp\left( -\frac{\delta_{k,\ell}}{u_{k,\ell}} \right),$$

- $IG(\gamma_k, \delta_{k,\ell})$ is the Inverse-Gamma distribution with scale parameter $\delta_{k,\ell}$ and shape parameter $\gamma_k$.

- We factorise the scale parameter $\delta_{k,\ell} = \omega_{fk} h_{k\ell}$.

- The NMF is placed on the hyperparameter, instead of $u_{k,\ell}$. 
Proposed Model Highlights

- Number of parameters: almost same with NMF;
- the $K$ additional $\gamma_k$ control the relevance of $u_{k,\ell}$.
- $u_{k,\ell}$ is of full rank $\Rightarrow |s_{j,\ell}|^2$ is of full rank;
- potentially allows more flexible demixing Wiener-filters;
Associated Graphical Model

\[ \gamma_k, w_{fk}, h_{k\ell} \rightarrow u_{k,f\ell} \rightarrow c_{f\ell} \rightarrow x_{f\ell} \rightarrow v_f, A_f \]
Inference & EM Algorithm

• Probabilistic inference of:

\[ C = \{c_{f\ell}\}_{f,\ell}, \ U = \{u_{k,f\ell}\}_{f,\ell,k} \text{ given } \mathcal{X} = \{x_{f\ell}\}_{f,\ell}. \]

• Gaussian sensor noise: \( p(\mathcal{X}|C) = \mathcal{N}_c( A_f G c_{f\ell}, v_f I ) \).

• A standard EM alternates between:

  • Inference of \( p(C, U|\mathcal{X}) \).
  • Estimation of \( \theta = \{v_f, w_{fk}, h_{k\ell}, A_f, \gamma_k\}_{f,\ell,k} \).

• Inference of \( p(C, U|\mathcal{X}) \) is intractable in our case;
Variational EM

- Variational approximation: $p(C,U|X) \approx p(C|X)p(U|X)$,
- E-step split into two steps:
  - Components E-step: Estimate $p(C|X)$ given $p(U|X)$
  - Component’s PSD E-step: Estimate $p(U|X)$ given $p(C|X)$.
Expectation Step - Components

- Components E-step: \( p(c_{f\ell}|X) = \mathcal{N}_c(\hat{c}_{f\ell}, \Sigma^c_{f\ell}) \) with

\[
\Sigma^c_{f\ell} = \left[ \text{diag}_K \left( \ldots, \frac{1}{\hat{u}_{k,f\ell}}, \ldots \right) + \frac{(A_f G)^H A_f G}{\nu_f} \right]^{-1},
\]

\(\hat{c}_{f\ell} = \Sigma^c_{f\ell} (A_f G)^H \frac{X_{f\ell}}{\nu_f} \).

- \( \hat{u}_{k,f\ell} \in \mathbb{R}_+ \) is given from the "old" \( p(U|X) \).

- The sources \( \hat{s}_{f\ell} \in \mathbb{C}^J \) are extracted with:

\[
\hat{s}_{f\ell} = G \hat{c}_{f\ell},
\]
Expectation Step - PSD (of components)

- Component’s PSD E-step:
  \[
  \hat{u}_{k,\ell} = \frac{\sum_{kk,\ell} c_{kk} + |\hat{c}_{k,\ell}|^2 + w_{fk} h_{k\ell}}{\gamma_k + 1}.
  \]

- \(\hat{u}_{k,\ell}\) is full rank!

- Increasing \(\gamma_k\) decreases the contribution of \(c_{k,\ell}\).
Maximization Step

- The parameter set $\theta = \{A_f, v_f, w_{fk}, h_{k\ell}, \gamma_k\}_{f, \ell, k}$ is updated by maximizing the complete data expected log-likelihood $\triangleq \mathbb{E}_{p(C|X)p(U|X)}[\log p(X, C, U)]$.

- LS estimators for $A_f$ and $v_f$;
- Updates for $w_{fk}, h_{k\ell}$: conceptually similar with IS-NMF$^4$.
- **scale-invariant update** for $\gamma_k$:

$$
\gamma_k = \frac{FL}{\sum_{f, \ell=1}^{F, L} \log \left( 1 + \frac{\Sigma_{kk, f\ell}^{c} + \hat{c}_{k, f\ell}}{w_{fk}h_{k\ell}} \right)}.
$$

$^4$C. Févotte, N. Bertin and J. L. Durrieu, 2009
Experimental Setup

- Convolutive stereo mixtures, 3 speech signals from TIMIT (length = 2s),
- Simulations using BRIR\[^5\] with $T_{60} = 680$ms.
- Comparison with NMF-MASS method\[^1\].
- Initialization of mixing matrices: \textbf{blind!} (the entries of $A_f$ set to 1). Initialization of NMF ($K_j = 20$): \textbf{corrupted} versions of the true source’s spectra:
- Performance evaluation using SDR\[^6\] (higher the better).

\[^5\]C. Hummersone, R. Mason and T. Brookes. 2013
\[^1\]Ozerov & Févotte 2010
\[^6\]E. Vincent, R. Gribonval, and C. Févotte, 2006
Quantitative Results

Average SDR (dB) scores on 10 sets of speakers:

<table>
<thead>
<tr>
<th>Corrupt.</th>
<th>Proposed</th>
<th>Baseline$^{[1]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$s_1$</td>
<td>$s_2$</td>
</tr>
<tr>
<td>20dB</td>
<td>8.6</td>
<td>6.2</td>
</tr>
<tr>
<td>10dB</td>
<td>8.3</td>
<td>6.0</td>
</tr>
<tr>
<td>0dB</td>
<td>2.6</td>
<td>1.7</td>
</tr>
</tbody>
</table>

SDR measured at the input:

<table>
<thead>
<tr>
<th>SDR(dB)</th>
<th>$s_1$</th>
<th>$s_2$</th>
<th>$s_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.3</td>
<td>-7.0</td>
<td>-2.7</td>
<td></td>
</tr>
</tbody>
</table>

$^{[1]}$[Ozerov & Févotte 2010]
Estimated Values of the Shape Parameter $\log(\gamma_k)$

Figure: High $\gamma_k \Rightarrow$ irrelevant component!
Conclusions and Future Work

- We propose an NMF "without factorisation" to parameterize $|s_{j, f\ell}|^2$, for MASS.
- Our model includes a component weighting mechanism.
- Results obtained with 3 sources and 2 microphones (underdetermined mixtures) are quite encouraging;
- We plan to thoroughly investigate initialization strategies to address blind setups.
Thank you!