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

gipsa-lab
UNIVERSITY OF TRENTO
Bar-Ilan University

Source Separation from Convolutive Mixtures

- 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:

$$\mathbf{x}_{f\ell} = \mathbf{A}_f \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}$$

mixture $[I \times 1]$
observed

mixing matrix $[I \times J]$
unknown!

source STFT $[J \times 1]$
unknown!

sensor noise $[I \times 1]$
unknown!

- $f = [1, F]$: frequency bins, $\ell = [1, L]$: time frames.

Outline of the General Methodology

- 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[1].
  - Estimate the mixing matrices $\mathbf{A}_f$.
  - Construct demixing Wiener Filters to extract $\mathbf{s}_{f\ell}$ from $\mathbf{x}_{f\ell}$.
  - Iterate ..

[1] [Ozerov and Févotte, 2010]

Local Gaussian Composite Model

- Inspired by[1][2]:
- Each source $s_{j,f\ell}$: sum of latent components

$$s_{j,f\ell} = \sum_{k=1}^{K_j} c_{k,f\ell} \Leftrightarrow \mathbf{s}_{f\ell} = \mathbf{G}\mathbf{c}_{f\ell},$$

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

- in total we have $K = \sum_{j=1}^{J} K_j$ components.
- Each component follows $p(c_{k,f\ell}) = \mathcal{N}_c\left(c_{k,f\ell}; 0, u_{k,f\ell}\right)$.

[1] [A. Ozerov and C. Févotte, 2010]
[2] [N. Q. K. Duong, E. Vincent and R. Gribonval, 2010]

Non-Negative Matrix Factorisation (NMF)

- Typically: $u_{k,f\ell} = w_{fk} h_{k\ell}$ as in[1][3]
- 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 flexiblity of the estimated demixing Wiener-filters due to low-rank constraint on $|s_{j,f\ell}|^2$.

[1] [A. Ozerov and C. Févotte, 2010]

[3] [S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, 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 $u_{k,f\ell} \in \mathbb{R}_+$ is considered as a r.v.

$$p(u_{k,f\ell}) = \mathcal{IG}\left(\ \gamma_k\ ,\ \delta_{k,f\ell}\ \right)$$

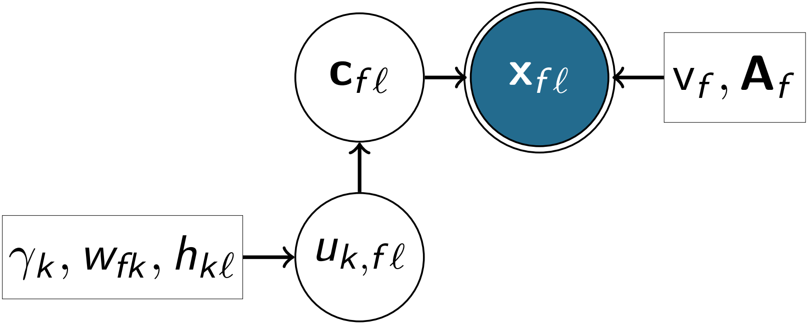
$$= \frac{(\delta_{k,f\ell})^{\gamma_k}}{\Gamma(\gamma_k)} {u_{k,f\ell}}^{-(\gamma_k+1)} \exp\left(-\frac{\delta_{k,f\ell}}{u_{k,f\ell}}\right),$$

- $\mathcal{IG}\left(\gamma_k, \delta_{k,f\ell}\right)$ is the Inverse-Gamma distribution with scale parameter $\delta_{k,f\ell}$ and shape parameter γ_k.
- **we factorise the scale parameter** $\delta_{k,f\ell} = w_{fk} h_{k\ell}$.
- The NMF is placed on the hyperparamter, instead of $u_{k,f\ell}$.

Proposed Model Highlights

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

Associated Graphical Model



Inference & EM Algorithm

- Probabilistic inference of:

  $\mathcal{C} = \{\mathbf{c}_{f\ell}\}_{f,\ell}, \mathcal{U} = \{u_{k,f\ell}\}_{f,\ell,k}$ given $\mathcal{X} = \{\mathbf{x}_{f\ell}\}_{f,\ell}$.

- Gaussian sensor noise: $p(\mathcal{X}|\mathcal{C}) = \mathcal{N}_c(\ \mathbf{A}_f \mathbf{G} \mathbf{c}_{f\ell},\ \mathrm{v}_f \mathbf{I}_I\ )$.
- A standard EM alternates between:
  - Inference of $p(\mathcal{C}, \mathcal{U}|\mathcal{X})$.
  - Estimation of $\theta = \left\{\mathrm{v}_f, w_{fk}, h_{k\ell}, \mathbf{A}_f, \gamma_k\right\}_{f,\ell,k}$.
- Inference of $p(\mathcal{C}, \mathcal{U}|\mathcal{X})$ is intractable in our case;

Variational EM

- Variational approximation: $p(\mathcal{C},\mathcal{U}|\mathcal{X}) \approx p(\mathcal{C}|\mathcal{X})p(\mathcal{U}|\mathcal{X})$,
- E-step split into two steps:
  - Components E-step: Estimate $p(\mathcal{C}|\mathcal{X})$ given $p(\mathcal{U}|\mathcal{X})$
  - Component's PSD E-step: Estimate $p(\mathcal{U}|\mathcal{X})$ given $p(\mathcal{C}|\mathcal{X})$.
- M-step: Estimation of $\mathbf{A}_f, \mathrm{v}_f$ and Inverse-Gamma parameters: via maximization of the complete-data expected log-likelihood.

Expectation Step - Components

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

$$\mathbf{\Sigma}^c_{f\ell} = \left[\mathrm{diag}_K\left(.., \frac{1}{\hat{u}_{k,f\ell}}, ..\right) + \frac{\left(\mathbf{A}_f\mathbf{G}\right)^{\mathrm{H}}\mathbf{A}_f\mathbf{G}}{\mathsf{v}_f}\right]^{-1},$$

$$\hat{c}_{f\ell} = \mathbf{\Sigma}^c_{f\ell}\left(\mathbf{A}_f\mathbf{G}\right)^{\mathrm{H}}\frac{\mathbf{x}_{f\ell}}{\mathsf{v}_f}.$$

- $\hat{u}_{k,f\ell} \in \mathbb{R}_+$ is given from the "old" $p(\mathcal{U}|\mathcal{X})$.
- The sources $\hat{\mathbf{s}}_{f\ell} \in \mathbb{C}^J$ are extracted with:

$$\hat{\mathbf{s}}_{f\ell} = \mathbf{G}\hat{\mathbf{c}}_{f\ell},$$

Expectation Step - PSD (of components)

- Component's PSD E-step:

$$\hat{u}_{k,f\ell} = \frac{\Sigma^c_{kk,f\ell} + |\hat{c}_{k,f\ell}|^2 + w_{fk} h_{k\ell}}{\gamma_k + 1}.$$

- $\hat{u}_{k,f\ell}$ is full rank!
- Increasing γ_k **decreases the contribution** of $c_{k,f\ell}$.

Maximization Step

- The parameter set $\theta = \{\mathbf{A}_f, \mathrm{v}_f, w_{fk}, h_{k\ell}, \gamma_k\}_{f,\ell,k}$ is updated by maximizing the complete data expected log-likelihood$\triangleq$

$$\mathbb{E}_{p(\mathcal{C}|\mathcal{X})p(\mathcal{U}|\mathcal{X})}\left[\log p(\mathcal{X}, \mathcal{C}, \mathcal{U})\right].$$

- LS estimators for $\mathbf{A}_f$ and v_f;
- Updates for $w_{fk}, h_{k\ell}$: conceptually similar with IS-NMF[4].
- **scale-invariant update** for γ_k:

$$\gamma_k = \frac{FL}{\sum_{f,\ell=1}^{F,L} \log\left(1 + \frac{\Sigma^c_{kk,f\ell} + |\hat{c}_{k,f\ell}|^2}{w_{fk} h_{k\ell}}\right)}.$$

[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: **blind!** (the entries of $\mathbf{A}_f$ set to 1). Initialization of NMF ($K_j = 20$): **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. Fevótte, 2006]

Quantitative Results

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

	Proposed			Baseline[1]		
Corrupt.	s_1	s_2	s_3	s_1	s_2	s_3
20dB	**8.6**	**6.2**	**9.3**	8.3	5.7	8.1
10dB	**8.3**	**6.0**	**8.0**	8.1	5.8	7.5
0dB	**2.6**	**1.7**	**0.8**	1.7	0.8	0.2

SDR measured at the input:

	s_1	s_2	s_3
SDR(dB)	-0.3	-7.0	-2.7

[1] [Ozerov & Févotte 2010]

Estimated Values of the Shape Parameter $\log(\gamma_k)$

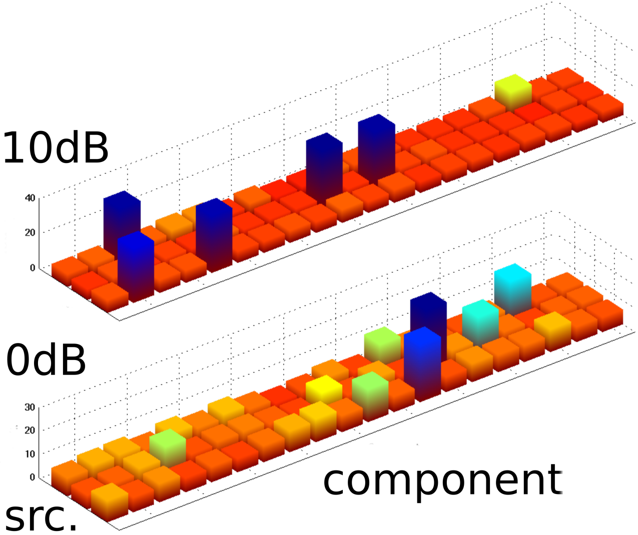


Figure: High $\gamma_k \Rightarrow$ irrelevant component!

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 !