Bayesian Hierarchical Modeling for Music and Audio Processing at LabROSA

Dawen Liang (LabROSA)

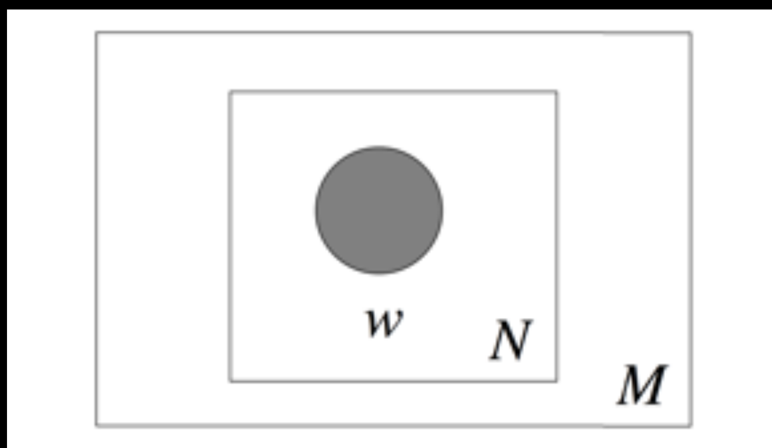
Joint work with: Dan Ellis (LabROSA), Matt Hoffman (Adobe Research), Gautham Mysore (Adobe Research)

1. Bayesian Hierarchical Modeling in general
2. Beta Process Sparse NMF
3. Product-of-Filters (PoF)

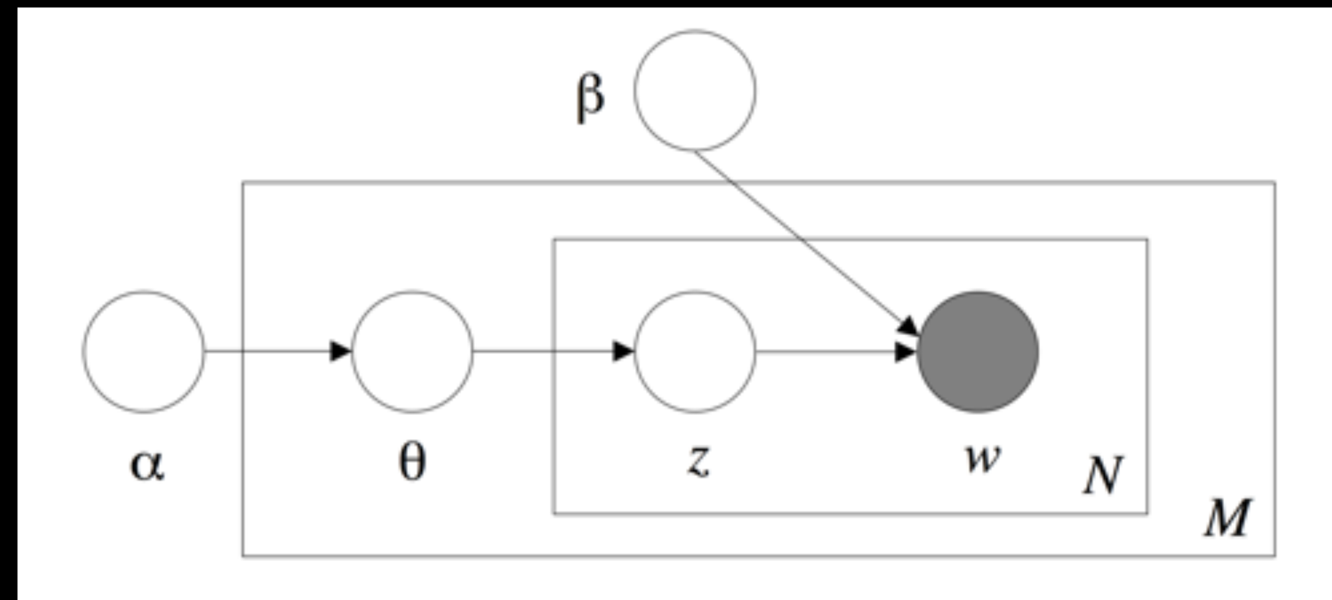
Lab ROSA

Laboratory for the Recognition and Organization of Speech and Audio

Bayesian Hierarchical Modeling in general



Unigram



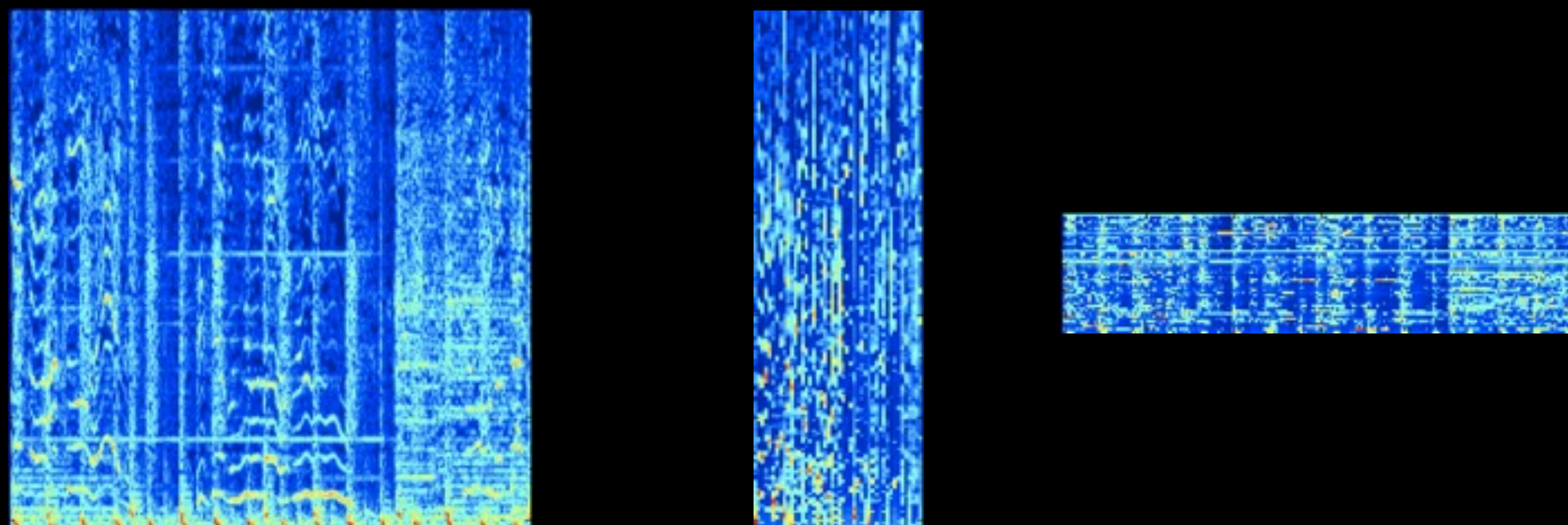
Latent Dirichlet Allocation

- Bayesian model with hierarchy
  - Bayesian model: interested in posterior $p(\Theta|X)$
  - Hierarchy: Latent variable layer(s)

NMF for Source Separation

(Lee & Seung, 2001; Smaragdis & Brown 2003)

$$X_{(F \times T)} \approx D_{(F \times K)} \; S_{(K \times T)}$$



D_{fk}: amplitude at freq f for source k

S_{kt}: gain of source k at time t

X_{ft}: observed amplitude at time t, freq f

Limitations of NMF

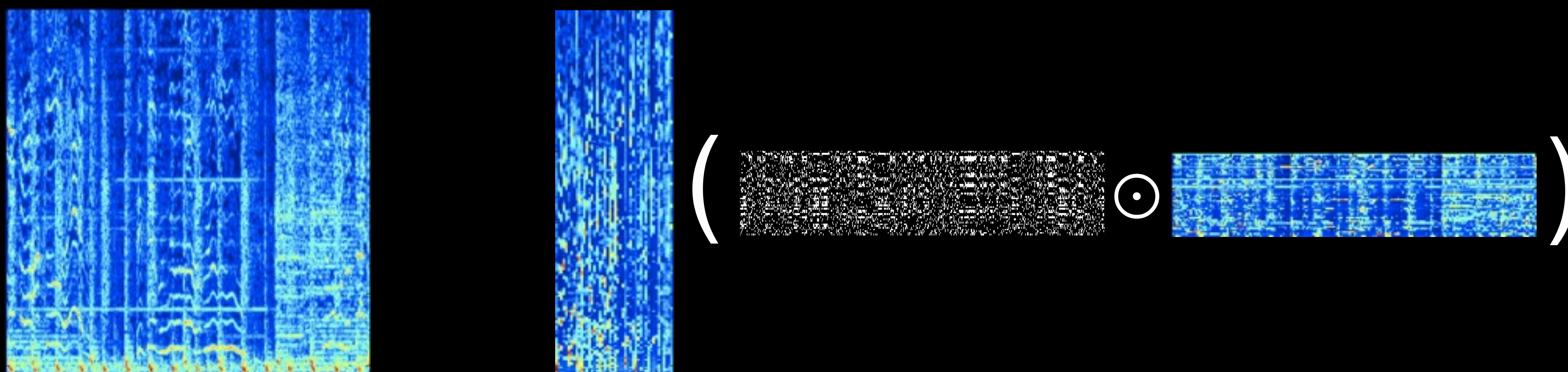
- It's not clear how to choose K, the number of components.
- The decomposition doesn't express the intuition that most components should be *silent* most of the time.

Beta Process Sparse NMF

- We propose Beta Process Sparse NMF (BP-NMF), a Bayesian nonparametric sparse NMF model.
  - Bayesian nonparametric: puts a *prior* on how many components will be used to explain the data.
  - Sparse: *explicitly silences* most components most of the time.

BP-NMF Decomposition

$$X_{(F \times T)} \approx D_{(F \times K)} \, (Z_{(K \times T)} \odot S_{(K \times T)})$$



D_{fk}: amplitude at freq f for source k

S_{kt}: gain of source k at time t

Z_{kt}: binary mask on source k at time t

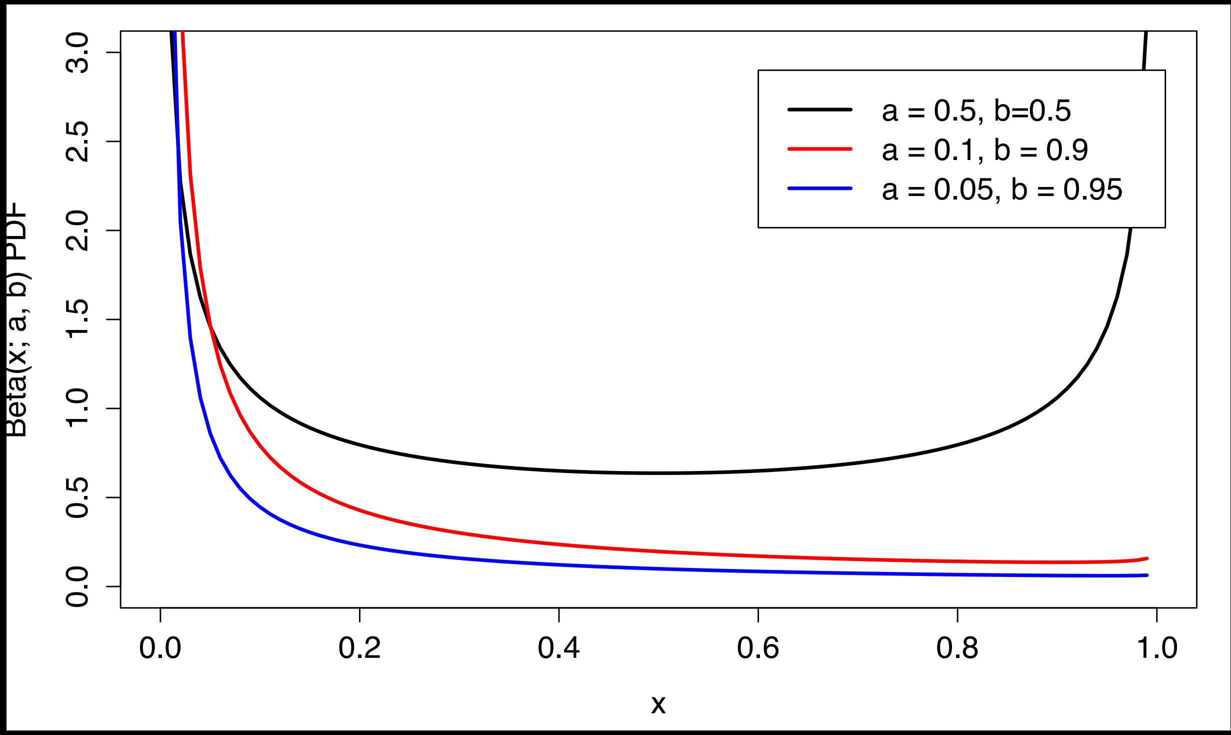
X_{ft}: observed amplitude at time t, freq f

The BP-NMF Model

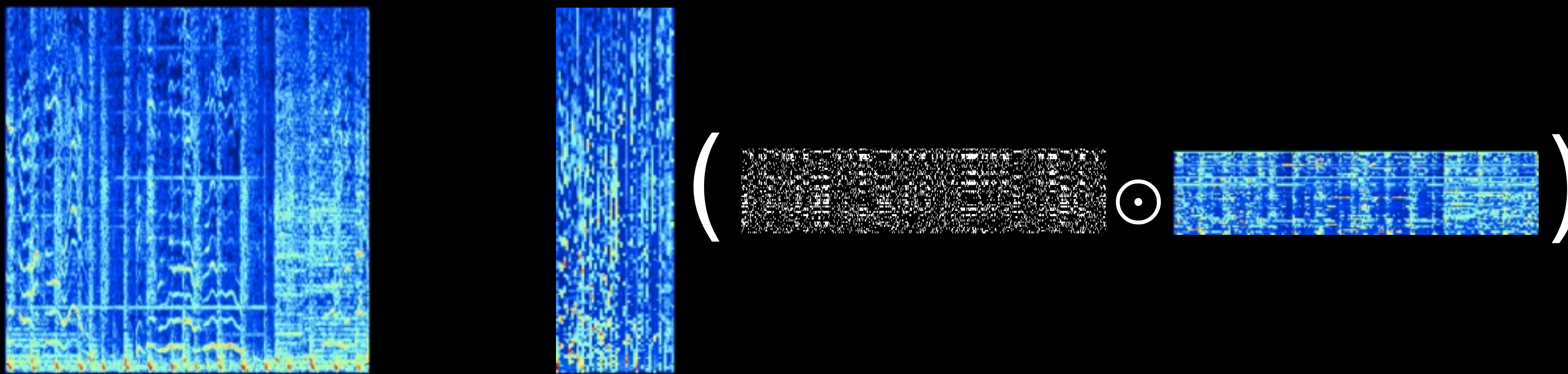
$$\pi_k \sim \text{Beta}(a_0/K, b_0(K-1)/K)$$
$$Z_{kt} \sim \text{Bernoulli}(\pi_k)$$

- Each source k has some probability π_k of being on at each time t.

- As K gets large, the expected number of elements of π that are significantly greater than 0 stays constant.



$$X_{(F \times T)} \approx D_{(F \times K)} \left(Z_{(K \times T)} \odot S_{(K \times T)}\right)$$



$$\log(\mathrm{D}_{fk}) \sim \mathrm{Normal}(0,1) \qquad \mathrm{S}_{kt} \sim \mathrm{Gamma}(\alpha, \beta)$$

$$\mathrm{X}_{ft} \sim \mathrm{Normal}(\textstyle\sum_k \mathrm{D}_{fk}\mathrm{Z}_{kt}\mathrm{S}_{kt}, \gamma^{-1}) \qquad \gamma \sim \mathrm{Gamma}(c_0, d_0)$$

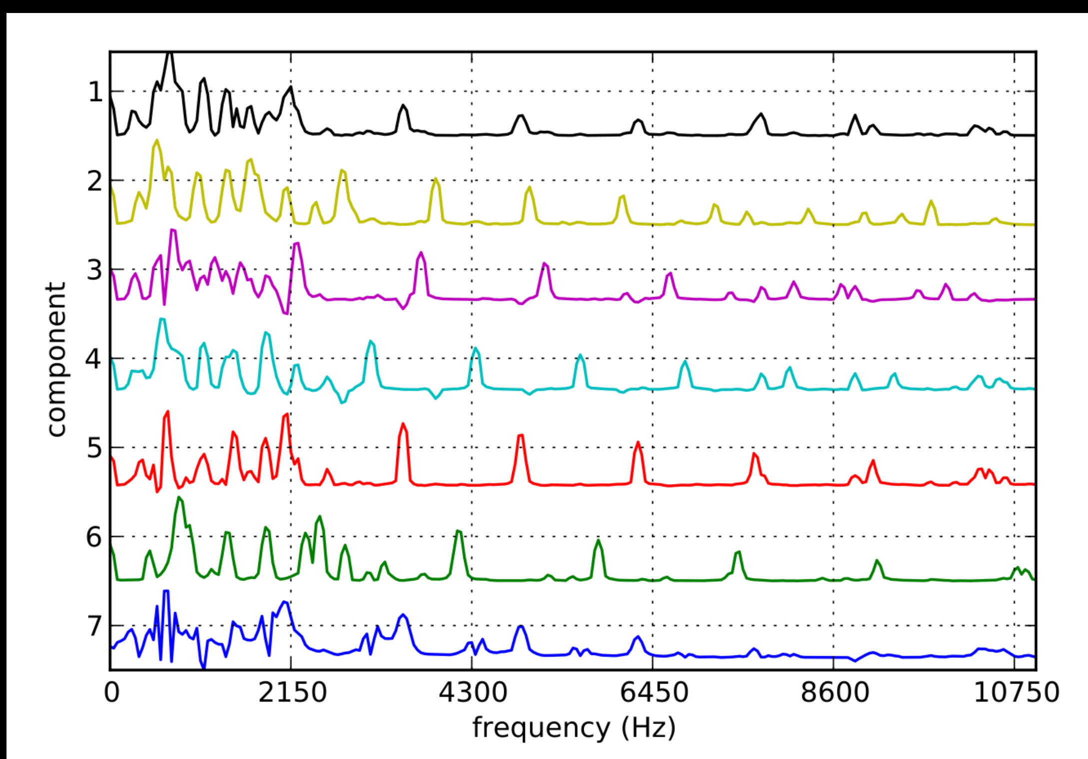
- Priors on **S**, **D**, and γ are chosen to preserve non-negativity, and for mathematical convenience.

Inference

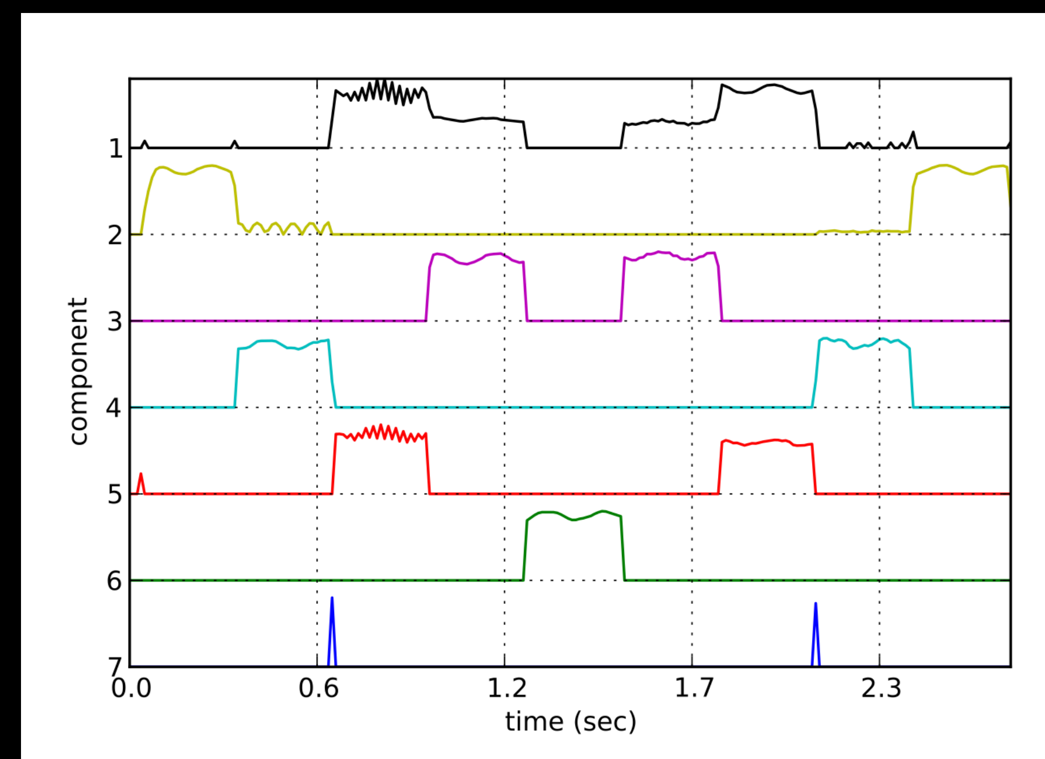
- The posterior is intractable, and the priors aren't conjugate.
- We use Laplace approximation variational inference (Wang & Blei, 2013) to approximate the posterior over the model parameters.
- Source code available (in Python!) at: https://github.com/dawenl/bp_nmf

Synthetic Data Experiment

- We ran BP-NMF on a synthetic recording of piano and clarinet.
  - One note from each instrument active at any given time.



Learned Dictionary **D**



Inferred Activations **S**⊙**Z**

Source Separation Experiment

- MIREX F_0 estimation data—woodwind quintet recording (bassoon, clarinet, flute, horn, oboe).
- We measure average BSS_eval metrics across each separated source, compare to GaP-NMF (Hoffman et al., 2010), another Bayesian nonparametric NMF model.
- BP-NMF discovers more components than GaP-NMF, which may be due to its ability to impose sparsity on the activations.

	SDR	SIR	SAR	K
BP-NMF	0.65	7.46	4.81	46
GaP-NMF	-1.86	3.89	6.12	31

Product-of-Filters (PoF)

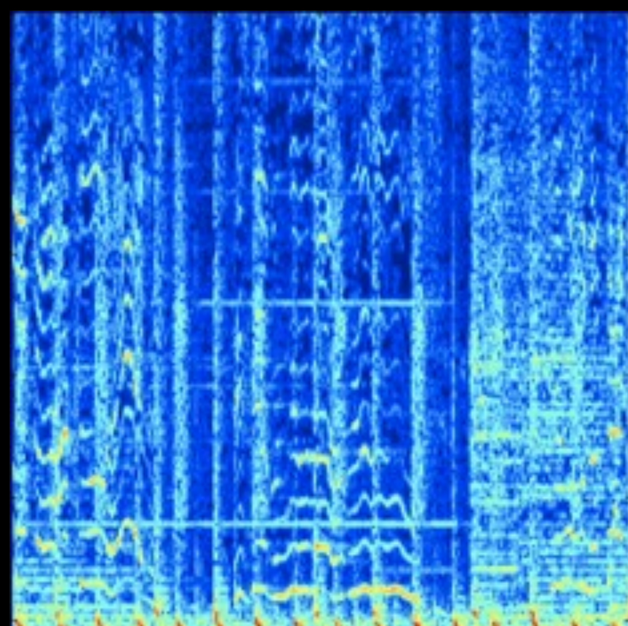
- Difference between NMF and PoF:
  - NMF decomposes *polyphonic* sounds into individual sources.
  - PoF decomposes *monophonic* sounds into simpler "*systems*" via statistical inference.

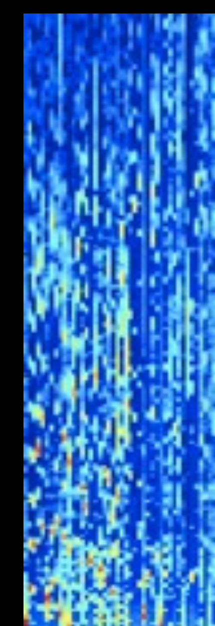
Motivation of PoF

- Homomorphic filtering:
    - A short window of audio is modeled as a convolution between an excitation signal and the impulse response of a series of linear filters.
- Likewise we model the observed magnitude spectra as a product of filters.

Product-of-Filters (PoF)

$$\log(W_{(F \times T)}) \approx U_{(F \times L)} \; A_{(L \times T)}$$

log ()



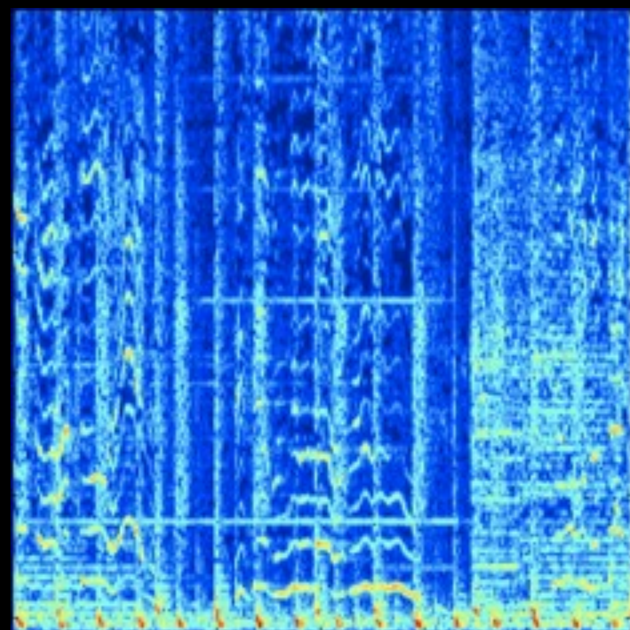
U_{fl}: amplitude on filter l at freq f

A_{lt}: activation on filter l at time t (non-negative)

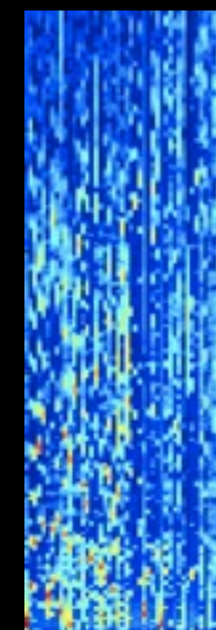
W_{ft}: observed amplitude at time t, freq f

$$\log W_{(F \times T)} \approx U_{(F \times L)} \; A_{(L \times T)}$$

log (



)

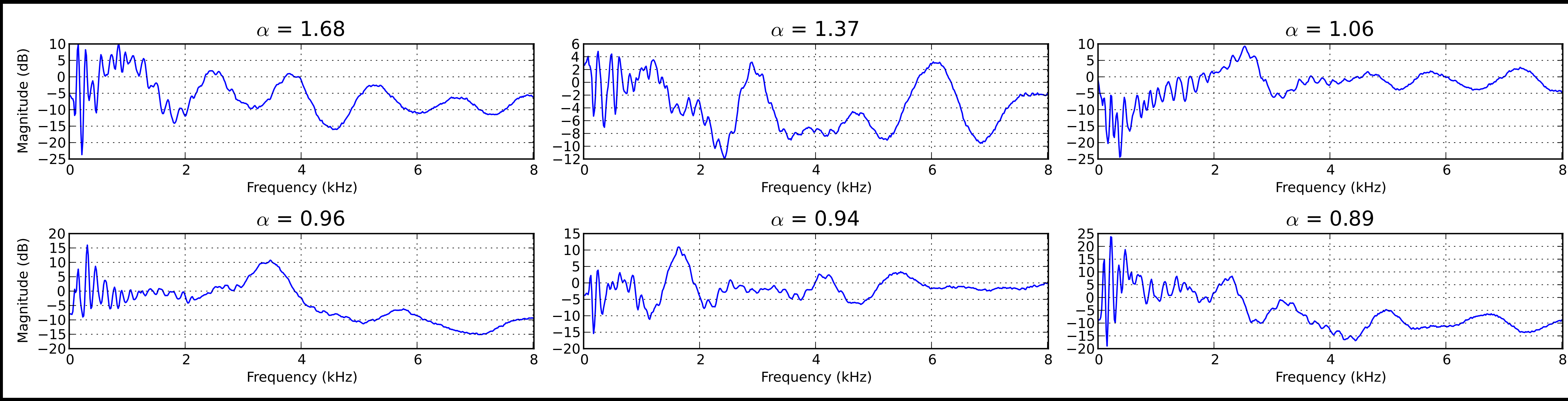


$$a_{lt} \sim \mathrm{Gamma}(\alpha_l, \alpha_l)$$
$$W_{ft} \sim \mathrm{Gamma}\Big(\gamma_f, \gamma_f / \exp(\textstyle\sum_l U_{fl} a_{lt})\Big)$$

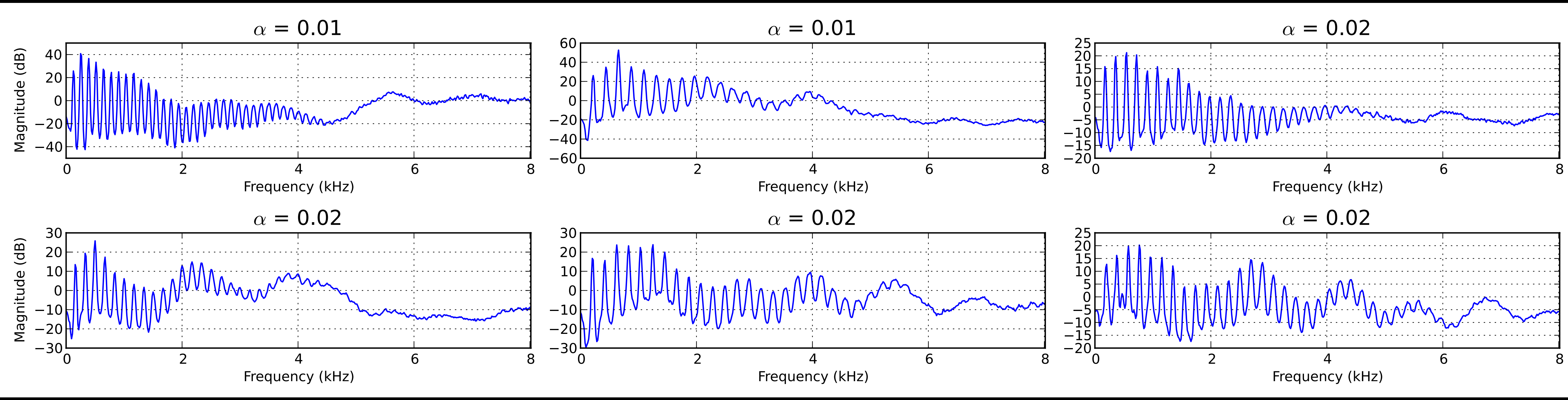
- Priors on **A** imposes sparsity, reflected by **α**, encoding the intuition that not all filters are always active.

- **A** is obtained via variational inference. **U**, γ, and **α** are learned via maximum marginal likelihood.

Filters Discovered from TIMIT



“vocal tract filter”



“excitation”

Experimental Results

- More details in the poster session
  - Bandwidth expansion
  - Speaker identification (v.s. MFCC)
  - (in progress) Phoneme classification with DNN

Wrapping up

- We briefly introduced Bayesian hierarchical modeling and two specific models for music and audio processing:
  - BP-NMF, a Bayesian nonparametric extension of the regular NMF model.
  - PoF, a novel model which decomposes *monophonic* sounds into simpler “*systems*” via statistical inference.

Questions?