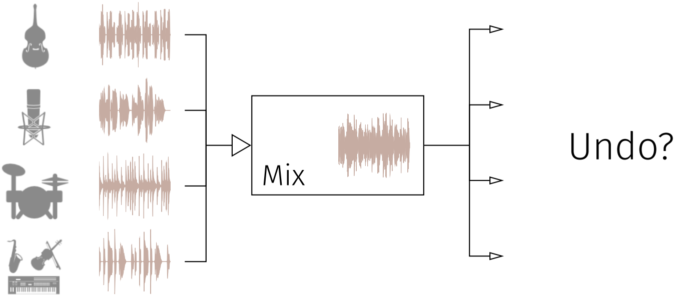
Unsupervised musical source separation

Maurice Frank

November 13, 2020

AMSTERDAM MACHINE LEARNING LAB
UNIVERSITY OF AMSTERDAM

The problem setup



Difference to the *cocktail party problem*:

Sources are different instruments with indivdual characteristics

The problem setup: notation

Source signals

$$\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]^T \in \mathbb{R}^{N \times T}$$

give the mix through the linear mixing function $f(\cdot)$

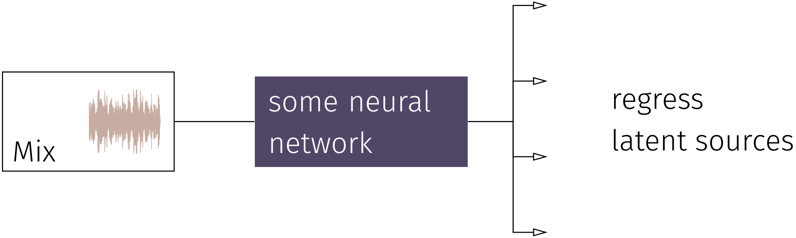
$$\boldsymbol{m} = f(\boldsymbol{S}) = \sum_k^N a_k \boldsymbol{s}_k \in \mathbb{R}^{1 \times T}$$

for which we search for an inverse $g(\cdot)$

$$g : \mathbb{R}^{1 \times T} \to \mathbb{R}^{N \times T}$$

$$g(\boldsymbol{m}) \approxeq \boldsymbol{S}$$

The common solution

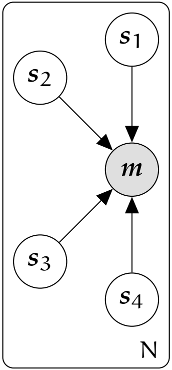


Optimize: $\arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N | \boldsymbol{m})$

Problem: Need expensive tuples $(\boldsymbol{m}, \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\})$

Bayesian perspective: Graphical model

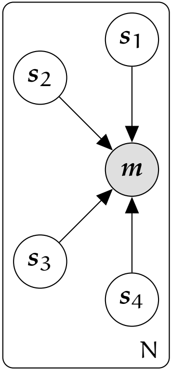
1. Do not learn separation ⇒ Learn generation $\mathrm{p}(\boldsymbol{m})$ instead
2. Source channels are latent variables and independent
3. Extract most likely set of latent values for a given mix $\boldsymbol{m}$



Bayesian perspective: Graphical model

Implied assumption

$$p(\boldsymbol{m}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) \equiv p(\boldsymbol{m}|\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) \cdot \prod_k^N p(\boldsymbol{s}_k)$$



that we can model the sources independently (*wrong!*).
How to retrieve get sources from the posterior?

$$\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N \sim p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N|\boldsymbol{m})$$

How to get sources from the posterior

With prior models $\{p_k(\boldsymbol{s}_k)\}_N$ how to get samples from the posterior

$$\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N \sim p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N | \boldsymbol{m})$$

- Sample from the posterior $\Rightarrow$ SGLD
- Model the posterior distribution $\Rightarrow$ VAE

Sampling from the posterior: SGLD

With *Stochastic Gradient Langevin Dynamics* we can directly sample from the posterior without modeling it.

Iteratively improve samples $\boldsymbol{s}_k$ with gradient update of the prior model and the mixing constraint.

$$\boldsymbol{s}_k^{(t+1)} = \boldsymbol{s}_k^{(t)} + \eta \cdot \nabla_{\boldsymbol{s}_k} \left(\log p_k(\boldsymbol{s}_k^{(t)}) + \frac{1}{2} \| \boldsymbol{m} - \sum_k^N \boldsymbol{s}_k^{(t)} \|^2 \right) + 2 \cdot \sqrt{\eta} \epsilon_t$$

Gaussian noise ϵ scaled by the step size η is added to avoid local maxima.

Jayaram & Thickstun, 2020 proved SGLD for separation for (small) images.

Propose an approximate posterior $\mathsf{q}_{\boldsymbol{\phi}_k}(\boldsymbol{s}_k|\boldsymbol{m})$ and we can derive the ELBO:

$$
\begin{aligned}
\mathbb{E}^{N}_{q_{\boldsymbol{\phi}_k}(\boldsymbol{s}_k|\boldsymbol{m})}\left[\log p(\boldsymbol{m})\right] &= \mathbb{E}^{N}_{q_{\boldsymbol{\phi}_k}(\boldsymbol{s}_k|\boldsymbol{m})}\left[\log \frac{p(\boldsymbol{m}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)}{p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N|\boldsymbol{m})}\right] \\
&\geq \sum_{k}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}_k}(\boldsymbol{s}_k|\boldsymbol{m})}\left[\log \frac{p(\boldsymbol{s}_k)}{q_{\boldsymbol{\phi}_k}(\boldsymbol{s}_k|\boldsymbol{m})}\right] \\
&\quad + \mathbb{E}_{q_{\boldsymbol{\phi}_k}(\boldsymbol{s}_k|\boldsymbol{m})}\left[p(\boldsymbol{m}|\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)\right]
\end{aligned}
$$

Modeling the posterior: VAE

Training of the encoders $\texttt{Encoder}_{\phi_k} : \boldsymbol{m} \to \mu, \sigma$ is done with KL-divergence and MSE loss.

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{m}) = \sum_{k}^{N} \mathbb{D}_{\mathrm{KL}} \left[q_{\boldsymbol{\phi}_k}(\boldsymbol{s}_k | \boldsymbol{m}) \| p_k(\boldsymbol{s}_k) \right] + \left(\frac{1}{N} \sum_{k}^{N} a_k \hat{\boldsymbol{s}}_k - \boldsymbol{m} \right)^2$$

$$\hat{\boldsymbol{s}}_k \sim q_{\boldsymbol{\phi}_k}(\boldsymbol{s}_k | \boldsymbol{m})$$

From theory to practice

musdb18

- 100 full real songs
- [bass, drums, voice, *other*]
- *No* post-mixing effects

ToyData

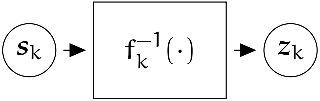
- Simple synthesizer waves
- [sine, saw, sqaure, triangle]
- Random phase, period and amplitude

Processing:

- Fixed-length frames, around 1sec
- Mix is simple mean

Modeling the priors

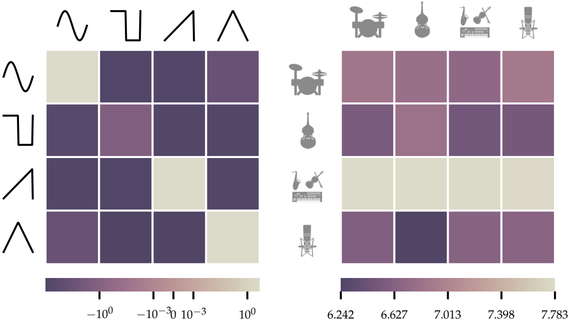
- Model $\log p_k(\boldsymbol{s}_k)$ with flow model.
- N Source priors are trained independently.
- Follow recent FloWaveNet[1]: Coupling layers are parameterized by WaveNets. Large receptive field through squeezing and dilated convolutions.



[1] S. Kim, S. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet : A Generative Flow for Raw Audio," May 2019

Experiment: Cross-likelihood

How likely are samples of one source type under another prior?



(a) ToyData **(b)** *musdb18*

Can we make at least work for the toy data?

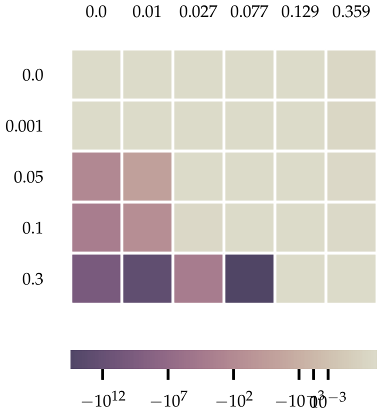
Noised prior distributions

Noiseless distribution is too spiked $\Rightarrow$ not useable for sampling or variational inference.

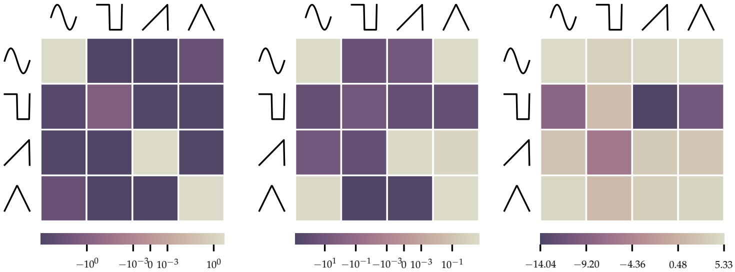
Fine-tune the noiseless models with increasing levels of noised examples to get noised distributions.

How likely are noisy examples under the nosied or noiseless prior distributions?

Here for the `sine` wave prior.

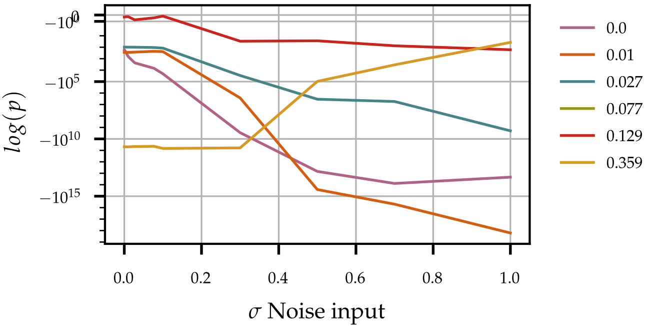


But, how discriminative are the priors for widening noise-conditioning?



(a) 0.0 **(b)** 0.027 **(c)** 0.129

Pure noise inputs show destruction of the noisy distribution:



Similar results for constant inputs 0 and 1:

value	model	sin	square	saw	triangle
0.0	0.0	4.8e+00	-7.0e+02	4.4e+00	1.8e+00
	0.359	-5.0e-01	-3.1e+00	5.1e+00	-2.0e+11
1.0	0.0	-1.5e+01	-3.6e+03	-2.7e+06	-3.2e+02
	0.359	2.7e+00	4.5e+00	-2.8e+00	-1.1e+01

Conclusion

1. New method for musical source separation without expensive training samples
2. Fails (so far) because priors are either not discriminative or not smooth enough
3. Confirm prior work on generative models and out-of-distribution samples