Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation

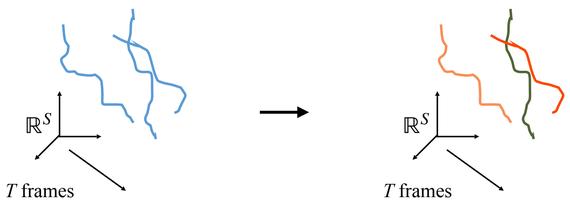
Xiaoyu Lin[1], Laurent Girin[2], Xavier Alameda-Pineda[1]

[1] Inria Grenoble Rhône-Alpes, Univ. Grenoble Alpes, France

[2] Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab, France

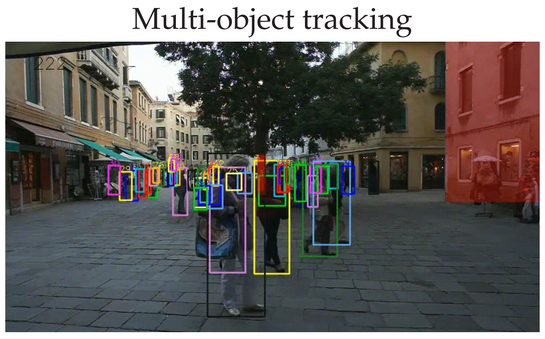
Context and Motivations

Separate multiple sources in sequential data

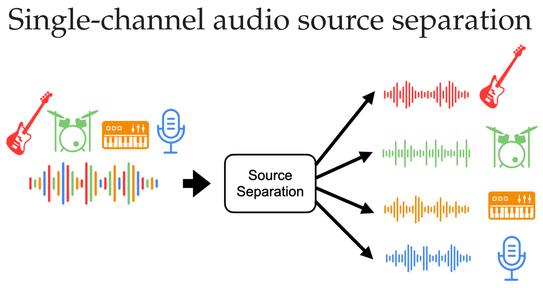


Application scenarios

Multi-object tracking



Single-channel audio source separation



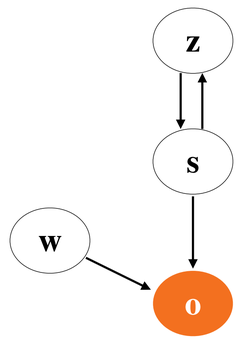
Limitations of supervised methods

- Necessity of extensive volumes of annotated data for training.
- Shortfall in model interpretability.
- Inadequate calibration of prediction uncertainty.

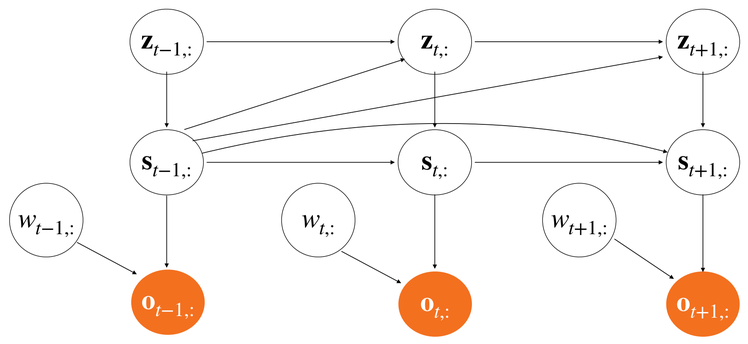
Proposed method

Leverage probabilistic generative models within an unsupervised or weakly supervised framework to solve the problem.

MixDVAE Probablistic Model



Folded graphical model



Extended graphical model over time frames

Definition of random variables

$\mathbf{o} \in \mathbb{R}^{T \times K_t \times O}$: observations.
$\mathbf{s} \in \mathbb{R}^{T \times N \times S}$: true source vectors.
$\mathbf{z} \in \mathbb{R}^{T \times N \times L}$: latent variables of DVAE.
$\mathbf{w} \in \{1, ..., N\}^{T \times K_t}$: assignment variables, $w_{tk} = n$ indicates observation $\mathbf{o}_{tk}$ assigned to source n.

Observed variable: $\mathbf{o}$ Latent variables: $\mathbf{s}$, $\mathbf{z}$, $\mathbf{w}$.

Objective: Estimate $p(\mathbf{s}, \mathbf{z}, \mathbf{w}|\mathbf{o})$.

Definition of probabilistic models

Generative model:

$$p_\theta(\mathbf{o}, \mathbf{w}, \mathbf{s}, \mathbf{z}) = p_{\theta_\mathbf{o}}(\mathbf{o}|\mathbf{w}, \mathbf{s}) p_{\theta_\mathbf{w}}(\mathbf{w}) p_{\theta_{\mathbf{sz}}}(\mathbf{s}, \mathbf{z})$$

Intractable true posterior distribution $p_\theta(\mathbf{s}, \mathbf{z}, \mathbf{w}|\mathbf{o})$

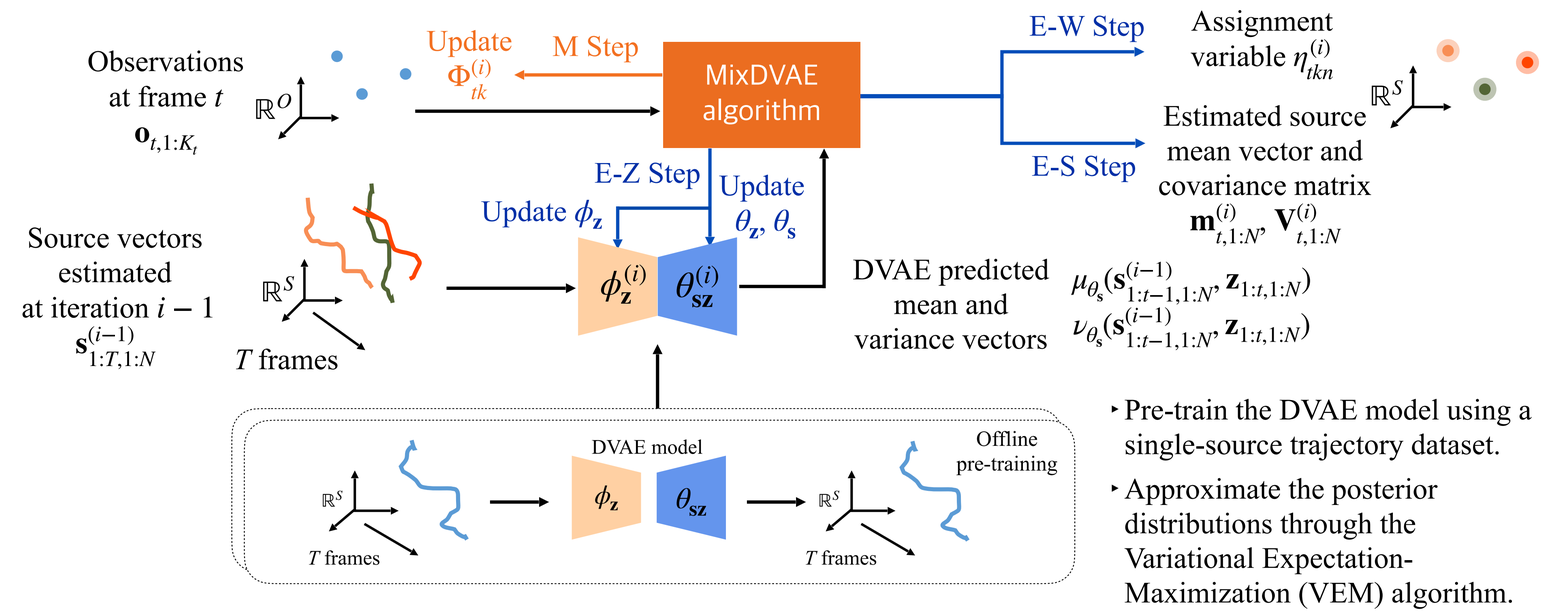
Inference model: mean-field like approximation

$$p_\theta(\mathbf{s}, \mathbf{z}, \mathbf{w}|\mathbf{o}) \approx q_{\phi_\mathbf{w}}(\mathbf{w}|\mathbf{o}) q_{\phi_\mathbf{z}}(\mathbf{z}|\mathbf{s}) q_{\phi_\mathbf{s}}(\mathbf{s}|\mathbf{o})$$

Optimization by maximizing the ELBO

$$\mathcal{L}(\theta, \phi; \mathbf{o}) = \mathbb{E}_{q_\phi(\mathbf{s},\mathbf{z},\mathbf{w}|\mathbf{o})}[\log p_\theta(\mathbf{o}, \mathbf{s}, \mathbf{z}, \mathbf{w}) - \log q_\phi(\mathbf{s}, \mathbf{z}, \mathbf{w}|\mathbf{o})].$$

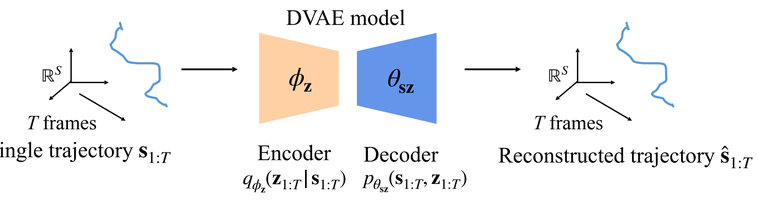
MixDVAE Algorithm



- Pre-train the DVAE model using a single-source trajectory dataset.
- Approximate the posterior distributions through the Variational Expectation-Maximization (VEM) algorithm.

Single-Source Dynamics Modeling

Dynamical Variational Autoencoders (DVAEs) are a family of powerful deep probabilistic generative models with latent variables designed for modeling sequential data with complex temporal dependencies [1]. The DVAE model is pre-trained on a synthetic or natural single-source dataset to embed prior knowledge about the complex data patterns.

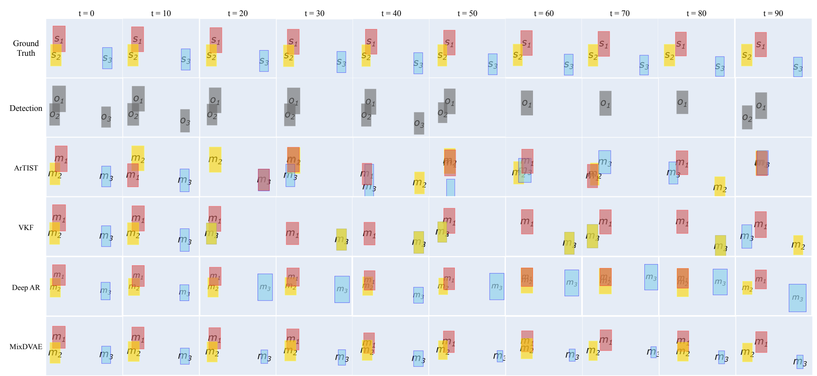


Pre-train DVAEs by maximizing the Evidence Lower BOund (ELBO):

$$\mathcal{L}(\theta, \phi; \mathbf{s}_{1:T}) = \mathbb{E}_{q_{\phi_\mathbf{z}}(\mathbf{z}_{1:T}|\mathbf{s}_{1:T})}[\log p_{\theta_\mathbf{sz}}(\mathbf{s}_{1:T}, \mathbf{z}_{1:T}) - \log q_{\phi_\mathbf{z}}(\mathbf{z}_{1:T}|\mathbf{s}_{1:T})]$$

Applications on MOT

MOT tracking example.

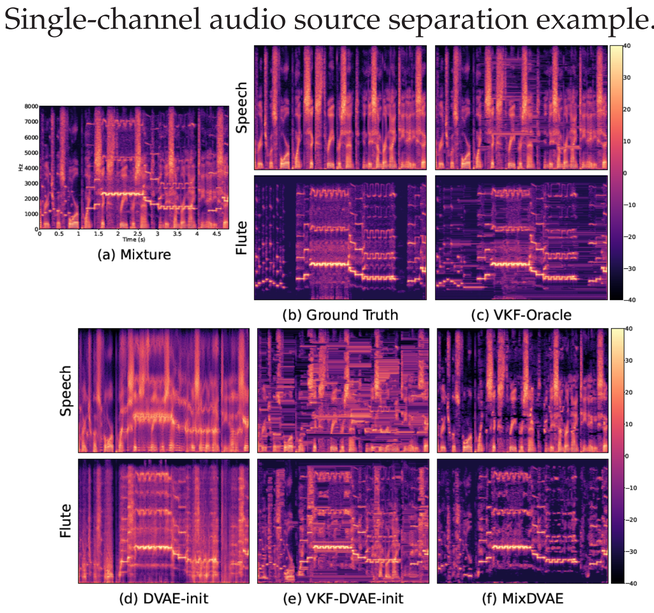


MOT results for short (T=60), medium (T=20), and long (T=300) sequences.

Dataset	Method	MOTA↑	MOTP↑	IDF1↑	#IDS↓	%IDS↓	MT↑	ML↓	#FP↓	%FP↓	#FN↓	%FN↓
Short	ArTIST	63.7	**84.1**	48.7	86371	28.0	**4684**	**0**	**9962**	**3.2**	**15525**	**5.0**
	VKF	56.0	82.7	77.3	5660	1.8	3742	761	64945	21.1	64945	21.1
	Deep AR	67.4	76.1	83.1	5248	1.7	3670	129	49595	16.0	49595	16.0
	MixDVAE	**79.1**	81.3	**88.4**	**4966**	**1.6**	4370	50	29808	9.7	29808	9.7
Medium	ArTIST	61.0	**84.2**	43.9	102978	24.6	**2943**	**0**	**25388**	**6.1**	**34812**	**8.3**
	VKF	57.5	83.3	77.6	7657	1.8	2563	487	85053	20.3	85053	20.3
	Deep AR	65.3	76.0	81.8	**5387**	**1.3**	2435	149	71775	17.0	71775	17.0
	MixDVAE	**78.6**	82.2	**88.0**	6107	1.5	2907	120	41747	9.9	41747	9.9
Long	ArTIST	53.5	84.5	40.7	205263	20.1	2513	**4**	135401	13.2	135401	13.2
	VKF	74.4	**86.2**	84.4	30069	2.9	2756	100	116160	11.4	116160	11.4
	Deep AR	75.5	76.6	87.1	26506	2.6	2555	18	123262	12.1	123262	12.1
	MixDVAE	**83.2**	82.4	**90.0**	**23081**	**2.3**	**2890**	12	**74550**	**7.3**	**74550**	**7.3**

Applications on SC-ASS

Single-channel audio source separation example.



SC-ASS results for short (T=50), medium (T=100), and long (T=300) sequences.

Dataset	Method	Speech RMSE ↓	Speech SI-SDR ↑	Speech PESQ ↑	Chinese bamboo flute RMSE ↓	Chinese bamboo flute SI-SDR ↑	Chinese bamboo flute PESQ ↑
Short	Mixture	0.016	-4.94	1.22	0.016	4.93	1.09
	VKF-Oracle	0.004	14.83	2.00	0.004	20.15	2.33
	DVAE-init	0.013	-0.51	1.20	0.019	3.04	1.44
	VKF-DVAE-init	0.012	2.24	1.21	0.012	8.06	1.33
	Deep AR	0.009	5.32	1.29	0.018	5.19	1.48
	MixIT	0.011	3.26	-	0.009	7.15	-
	Vanilla NMF	0.011	3.01	1.40	0.012	9.09	1.37
	Temporal NMF	0.009	4.99	1.53	0.011	10.26	1.53
	MixDVAE	**0.006**	**9.23**	**1.73**	**0.007**	**13.50**	**2.30**
Medium	Mixture	0.016	-4.44	1.17	0.016	4.44	1.08
	VKF-Oracle	0.004	14.88	1.88	0.003	20.24	2.41
	DVAE-init	0.014	0.10	1.15	0.020	2.42	1.27
	VKF-DVAE-init	0.013	1.25	1.12	0.013	7.42	1.26
	Deep AR	0.010	4.88	1.21	0.017	5.17	1.35
	MixIT	0.009	4.75	-	0.009	8.74	-
	Vanilla NMF	0.011	3.28	1.41	0.011	8.88	1.35
	Temporal NMF	0.010	5.12	1.48	0.011	9.96	1.44
	MixDVAE	**0.007**	**9.32**	**1.65**	**0.007**	**13.05**	**2.16**
Long	Mixture	0.016	-4.52	1.19	0.016	4.53	1.10
	VKF-Oracle	0.004	14.65	1.89	0.003	20.45	2.60
	DVAE-init	0.013	0.20	1.15	0.020	2.29	1.22
	VKF-DVAE-init	0.013	0.34	1.10	0.013	7.35	1.24
	Deep AR	0.010	3.87	1.17	0.017	4.74	1.27
	MixIT	**0.006**	**10.2**	-	0.007	11.76	-
	Vanilla NMF	0.011	3.31	1.40	0.011	8.98	1.35
	Temporal NMF	0.010	5.01	1.47	0.011	10.06	1.42
	MixDVAE	0.007	9.06	**1.64**	**0.007**	**12.92**	**2.06**

References

[1] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *Found. Trends Mach. Learn.*, 15(1-2):1–175, 2021.