Managing Multimodal Missingness

Factorized Inference in Deep Markov Models for Incomplete Multimodal Time Series

Tan Zhi-Xuan[1,2], Harold Soh[3], Desmond C. Ong[2,3]

[1]MIT EECS [2]A*STAR AI Initiative [3]NUS School of Computing

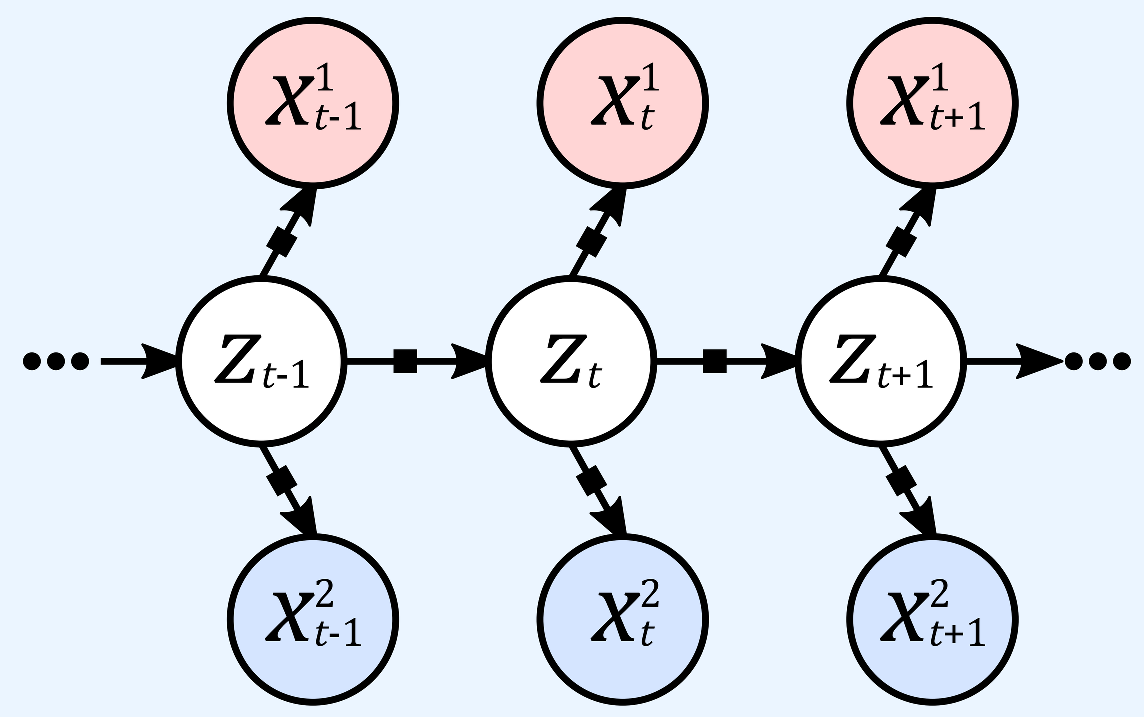
Motivation

- Incomplete multimodal time series data is highly common
  - mobile robots with asynchronous sensors
  - partially annotated videos for semantic segmentation
- Classical models (e.g. HMMs) are insufficiently powerful
- Neural models (e.g. RNNs) do not handle missingness
- Hybrid deep probabilistic models still rely on RNNs for inference

Contributions

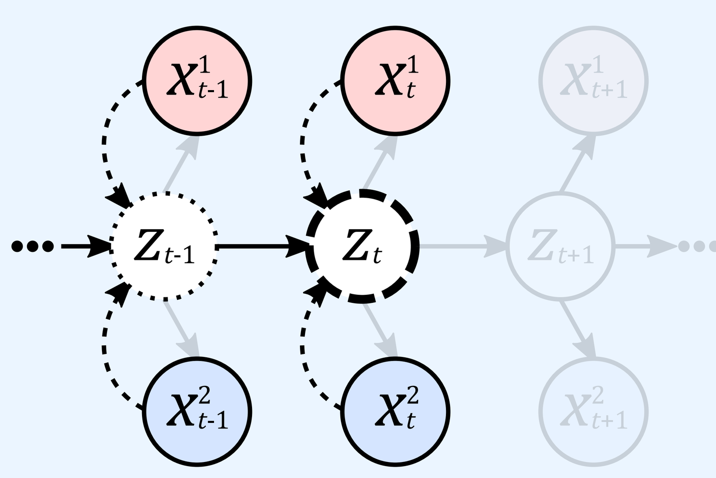
- A novel inference method that handles *multimodality* and *missingness*
- Combines strengths of *neural networks* with *message passing*
- Capable of *filtering*, *smoothing*, and *sequencing*
- Performs *interpolation*, *extrapolation*, and *conditional generation*
- Allows for *weakly supervised learning* of time series data

Multimodal Deep Markov Model (MDMM)



- Nonlinear Gaussian state space model
- Conditionally independent modalities $x^1, \ldots, x^M$
- Transitions and emissions modeled by deep neural networks

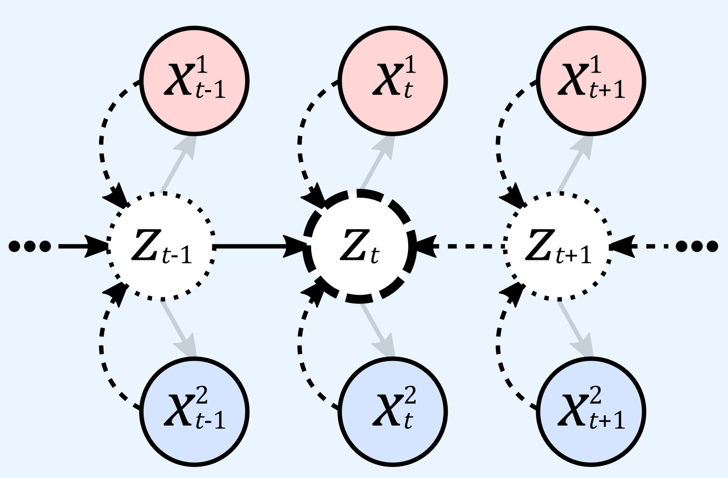
Factorized Inference in MDMMs



Filtering

Infer current z_t given *past* observations

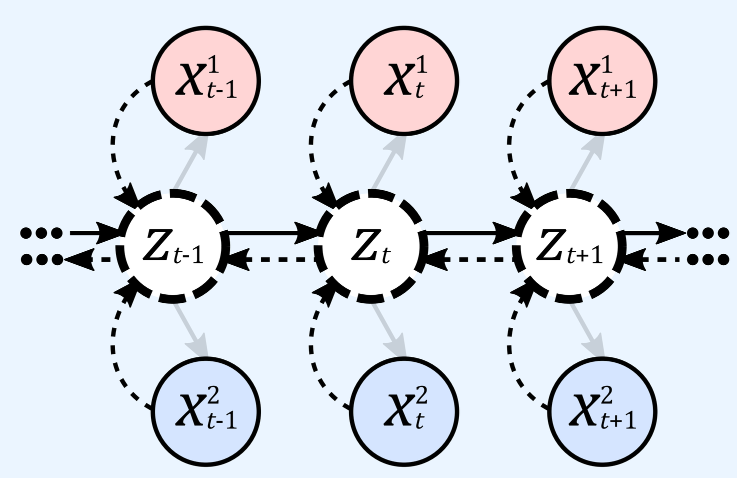
$$p(z_t|x_{1:t}) \propto \underbrace{p(z_t|z_{t-1})}_{\text{past}} \underbrace{\left[\prod_{m=1}^{M} \frac{p(z_t|x_t^m)}{p(z_t)}\right]}_{\text{present}}$$



Smoothing

Infer some z_t given *all* observations

$$p(z_t|x_{1:T}) \propto \underbrace{p(z_t|x_{1:t-1})}_{\text{past}} \underbrace{\left[\prod_{m=1}^{M} \frac{p(z_t|x_t^m)}{p(z_t)}\right]}_{\text{present}} \underbrace{\frac{p(z_t|x_{t+1:T})}{p(z_t)}}_{\text{future}}$$



Sequencing

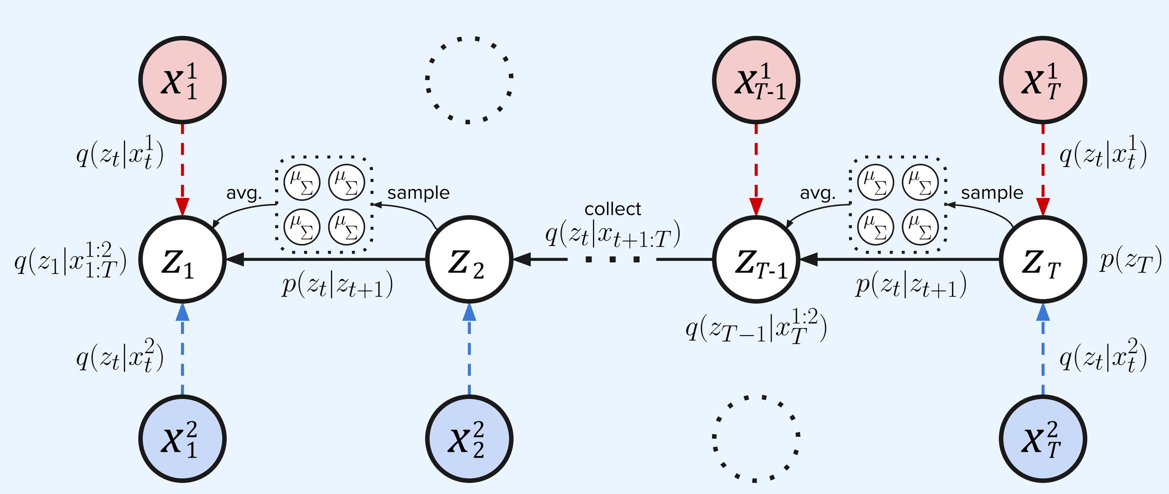
Infer z_1 to z_T given *all* observations

$$p(z_{1:T}|x_{1:T}) \propto \prod_{t=1}^{T} \underbrace{p(z_t|z_{t-1})}_{\text{past}} \underbrace{\left[\prod_{m=1}^{M} \frac{p(z_t|x_t^m)}{p(z_t)}\right]}_{\text{present}} \underbrace{\frac{p(z_t|x_{t+1:T})}{p(z_t)}}_{\text{future}}$$

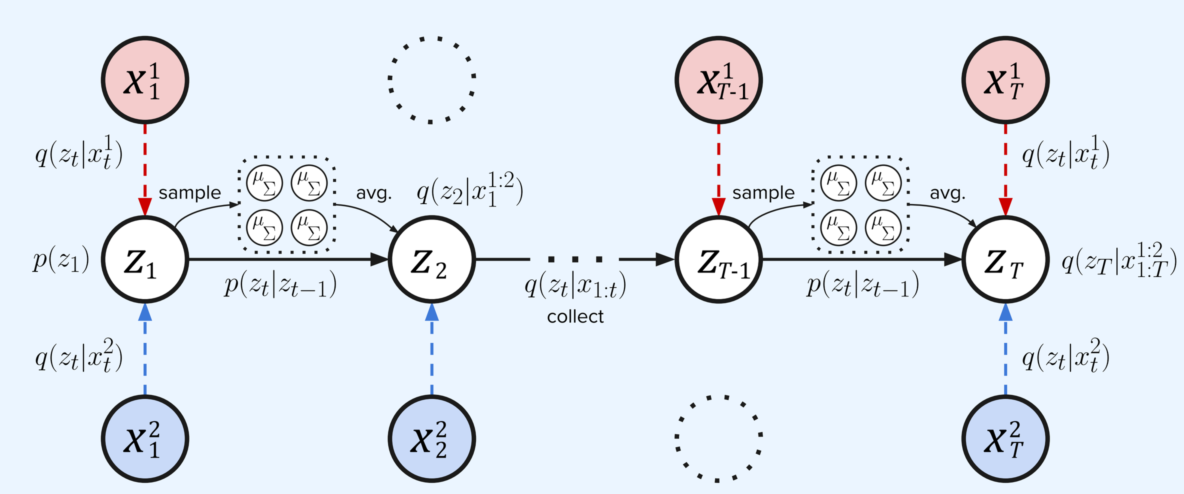
- Posteriors can be factorized into:
  - dependence upon past + present + future
  - dependence upon each modality
- Multimodal temporal fusion via:
  - approximating each term as Gaussian
  - multiplying via Product of Gaussians
- Advantages of Product of Gaussians:
  - tractable (weighted sum of input parameters)
  - handles missing modalities
  - gives more weight to more certain modalities
- Compute past and future dependence via:
  - backward message passing from the future
  - forward message passing from the past

Backward-Forward Variational Inference (BFVI)

Step 1: Backward pass to compute $q(z_t|x_{t+1:T})$



Step 2: Forward pass to compute $q(z_t|x_{1:t})$



Step 3: Combine forward and backward messages for smoothing or sequencing

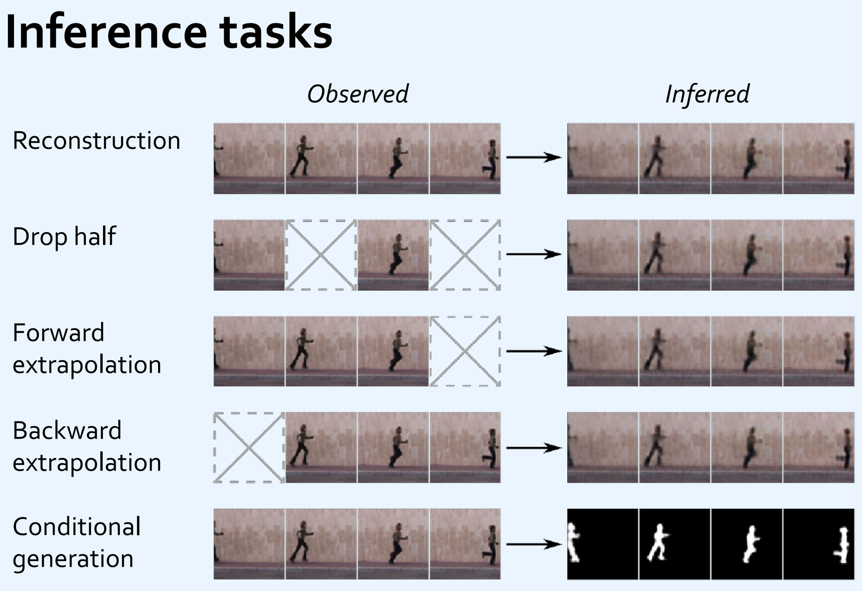
$$q(z_t|x_{1:T}) \propto q(z_t|x_{1:t})q(z_t|x_{t+1:T})/p(z_t)$$

Step 4: Maximize ELBO by learning neural network parameters for the model and inference distributions

Model:	$p(z_t|z_{t-1})$	$p(x_t^m|z_t)$
Inference:	$p(z_t|z_{t+1})$	$q(z_t|x_t^m)$

Experiments

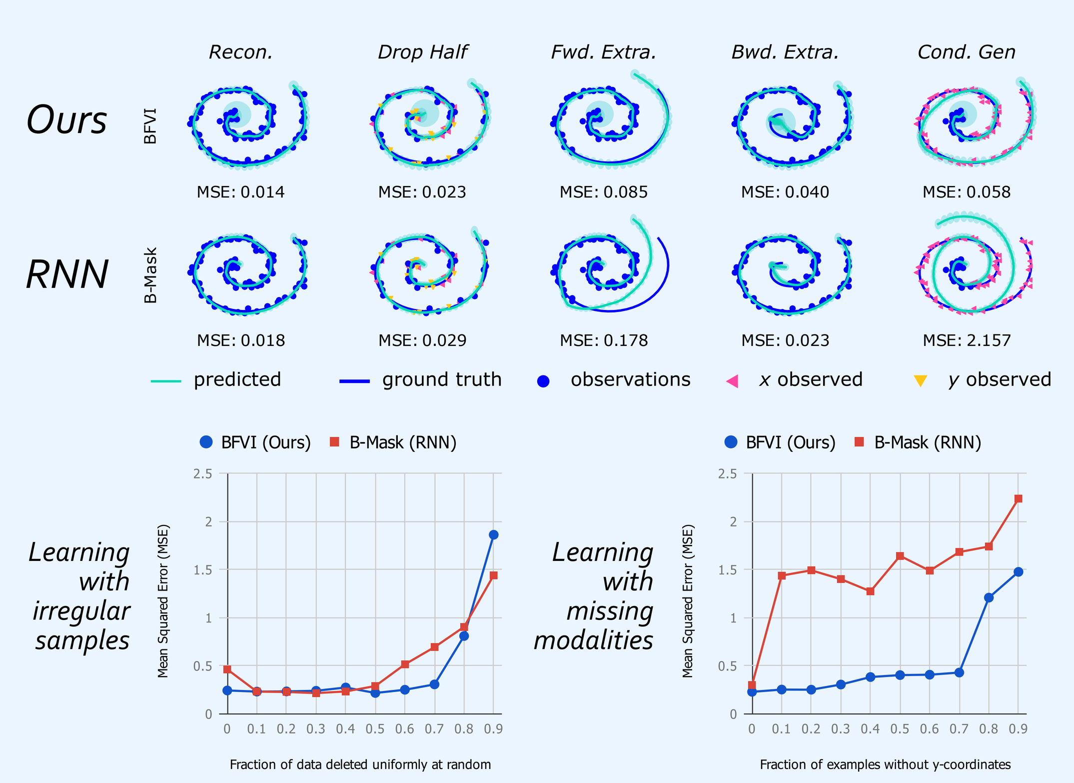
Inference tasks



Baselines

- Forward RNN to infer z_t given $x_{1:t}$
- Backward RNN to infer z_t given $x_{t:T}$
- Zero-masking / update-skipping variants

Dataset I: Noisy Spirals (x & y co-ordinates)



Dataset II: Weizmann actions (video + silhouettes + labels)

Email: xuan@mit.edu Code: https://git.io/Jeoze