Tan et al. 2019: Factorized Inference in Deep Markov Models for Incomplete Multimodal Time Series

Minqi Pan

March 17, 2020

Factorized Inference in Deep Markov Models for Incomplete Multimodal Time Series

- AAAI 2020 “ML: Probabilistic Methods II”, Feb 12nd, 2020
- Tan Zhi-Xuan, Harold Soh, Desmond C. Ong
- A*STAR, MIT, National University of Singapore

Outline

Outline

Multimodal Deep Markov Models (MDMMs)

- z_t: vector valued latent state
- x_t^m: vector valued observation for modality m at time t
- Define an MDMM with M modalities by
  - Transition distributions are assumed to be a multivariate Guassian with means and covariances which are differentiable functions of the previous latent state

$$z_t \sim \mathcal{N}(\mu_\theta(z_{t-1}), \Sigma_\theta(z_{t-1}))$$

  - Emission distributions

$$x_t^m \sim \Pi(\kappa_\theta^m(z_t))$$

    E.g. if the data is binary, $\Pi =$independent Bernoulli parameterized by $\kappa_\theta^m(z_t)$

Subsuming Linear Gaussian State Space Models

- $z_t \sim \mathcal{N}(\mu_\theta(z_{t-1}), \Sigma_\theta(z_{t-1}))$
- $x_t^m \sim \Pi(\kappa_\theta^m(z_t))$
- Kalman filters
  - $\mu_\theta(z_{t-1}) = G_t z_{t-1} + B_t u_t$ where G_t, B_t are a matrices
  - $\Sigma_\theta(z_{t-1}) = K_t$ where K_t is a matrix
  - $\kappa_\theta^m(z_t) = F_t z_t$ where F_t is a matrix
  - $\Pi = \mathcal{N}$
  - We can do inference analytically!
- Deep nonlinear models
  - $\mu_\theta(z_{t-1})$ is a neural network parameterized by θ
  - $\Sigma_\theta(z_{t-1})$ is a neural network parameterized by θ
  - $\kappa_\theta^m(z_t)$ is a neural network parameterized by θ

Jointly Learning θ (Generative) and ϕ (Inference)

- θ of the generative model $p_\theta(z_{1:T}, x_{1:T})$
  - ASSUMPTION: we consider learning in a Bayesian network whose joint distribution (generatively) factorizes as

$$p_\theta(z_{1:T}, x_{1:T}) = p_\theta(x_{1:T}|z_{1:T})p_\theta(z_{1:T})$$

  - Note that the marginal data likelihood is intractable:

$$p_\theta(x_{1:T}) = \int p_\theta(z_{1:T})p_\theta(x_{1:T}|z_{1:T})dz$$

- ϕ of the variational posterior $q_\phi(z_{1:T}|x_{1:T})$
  - $q_\phi(z_{1:T}|x_{1:T})$ approximates the true posterior $p_\theta(z_{1:T}|x_{1:T})$
  - $p_\theta(z_{1:T}|x_{1:T}) = \frac{p_\theta(x_{1:T}|z_{1:T})p_\theta(z_{1:T})}{p_\theta(x_{1:T})}$ is intractable

Evidence Lower Bound (ELBO)

$$\begin{aligned} L(x;\theta,\phi) =& \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}[\log p_\theta(x_{1:T}|z_{1:T})] \\ & - \mathbb{E}_{q_\phi(z_{1:T}|x_{1:T})}[\mathsf{KL}(q_\phi(z_{1:T}|x_{1:T})\|p_\theta(z_{1:T}))] \end{aligned}$$

- Jensen's inequality: L is a lower bound of the log marginal likelihood $L(x;\theta,\phi) \leqslant p_\theta(x_{1:T})$
- ML Learning $\Rightarrow$ Let's maximize L (via gradient ascent with stochastic backpropagation, sampling from q_ϕ)
- The expectation wrt $q_\phi(z_{1:T}|x_{1:T})$ implicitly depends on the network parameters ϕ. When using a Gaussian variational approximation $q_\phi(z_{1:T}|x_{1:T}) \sim \mathcal{N}(\mu_\phi(x_{1:T}), \Sigma_\phi(x_{1:T}))$, μ_ϕ, Σ_ϕ are parameteric functions of the observation

MDMMs can do 3 Kinds of Inferences

1. Filtering: given PAST, infer

$$p(z_t|x_{1:t}) \text{ for some } z_t$$

2. Smoothing: given PAST and FUTURE, infer

$$p(z_t|x_{1:T}) \text{ for some } z_t$$

3. Sequencing: given PAST and FUTURE, infer

$$p(z_{1:T}|x_{1:T})$$

Factorization over Time

$$
\begin{aligned}
p(z_{1:T}|x_{1:T}) &= p(z_1|x_{1:T})p(z_2|z_1, x_{1:T})p(z_3|z_2, x_{1:T})\dots \\
&= p(z_1|x_{1:T})p(z_2|z_1, x_{2:T})p(z_3|z_2, x_{3:T})\dots \\
&= p(z_1|x_{1:T})\prod_{t=2}^{T} p(z_t|z_{t-1}, x_{t:T})
\end{aligned}
$$

- Each latent state z_t depends only on
  - the previous latent state z_{t-1}
  - all current and future observations $x_{t:T}$

"Conditional Smoothing Posterior"

$$p(z_t|z_{t-1}, x_{t:T})$$

- it is the posterior that corresponds to the conditional prior $p(z_t|z_{t-1})$, hence we call it conditional "posterior"
- it combines information from both PAST and FUTURE, hence we call it "smoothing"

Factorizing the Conditional Smoothing Posterior (1)

- $x_{t:T}^{1:M} \perp\!\!\!\perp z_{t-1} | z_t$ (by d-seperation)

$$
\begin{aligned}
\Rightarrow p(z_t | z_{t-1}, x_{t:T}^{1:M}) &= \frac{p(z_{t-1}, z_t, x_{t:T}^{1:M})}{p(z_{t-1}, x_{t:T}^{1:M})} \\
&= \frac{p(x_{t:T}^{1:M} | z_{t-1}, z_t) p(z_{t-1}, z_t)}{p(z_{t-1}, x_{t:T}^{1:M})} \\
&= \frac{p(z_{t-1}) p(z_t | z_{t-1}) p(x_{t:T}^{1:M} | z_t)}{p(x_{t:T}^{1:M} | z_{t-1}) p(z_{t-1})}
\end{aligned}
$$

Factorizing the Conditional Smoothing Posterior (2)

- $x_t \perp\!\!\!\perp x_{t+1:T} | z_t$ (by Local Markov Property)

$$\begin{aligned}
\Rightarrow p(z_t|z_{t-1}, x_{t:T}^{1:M}) &= \frac{p(z_{t-1})p(z_t|z_{t-1})p(x_{t:T}^{1:M}|z_t)}{p(x_{t:T}^{1:M}|z_{t-1})p(z_{t-1})} \\
&= \frac{p(z_{t-1})p(z_t|z_{t-1})p(x_t^{1:M}|z_t)p(x_{t+1:T}^{1:M}|z_t)}{p(x_{t:T}^{1:M}|z_{t-1})p(z_{t-1})} \\
&= \frac{p(z_t|z_{t-1})p(x_t^{1:M}|z_t)p(x_{t+1:T}^{1:M}|z_t)}{p(x_{t:T}^{1:M}|z_{t-1})} \\
&= p(x_{t+1:T}^{1:M}|z_t)p(x_t^{1:M}|z_t)\frac{p(z_t|z_{t-1})}{p(x_{t:T}^{1:M}|z_{t-1})}
\end{aligned}$$

Factorizing the Conditional Smoothing Posterior (3)

- Dropping $\frac{1}{p(x_{t:T}^{1:M}|z_{t-1})}$
- Assuming $p(x_t^{1:M}|z_t) = \prod_{m=1}^{M} p(x_t^m|z_t)$

$$
\begin{aligned}
\Rightarrow p(z_t|z_{t-1}, x_{t:T}^{1:M}) &= p(x_{t+1:T}^{1:M}|z_t) p(x_t^{1:M}|z_t) \frac{p(z_t|z_{t-1})}{p(x_{t:T}^{1:M}|z_{t-1})} \\
&\propto p(x_{t+1:T}^{1:M}|z_t) p(x_t^{1:M}|z_t) p(z_t|z_{t-1}) \\
&= p(x_{t+1:T}^{1:M}|z_t) \left[\prod_{m=1}^{M} p(x_t^m|z_t) \right] p(z_t|z_{t-1})
\end{aligned}
$$

Factorizing the Conditional Smoothing Posterior (4)

- Dropping $p(x_{t+1:T}^{1:M}) \prod_{m=1}^{M} p(x_t^m) = p(x_{t:T}^{1:M})$

$$
\begin{aligned}
\Rightarrow\ & p(z_t | z_{t-1}, x_{t:T}^{1:M}) \\
&\propto p(x_{t+1:T}^{1:M} | z_t) \left[\prod_{m=1}^{M} p(x_t^m | z_t)\right] p(z_t | z_{t-1}) \\
&= \frac{p(z_t | x_{t+1:T}^{1:M}) p(x_{t+1:T}^{1:M})}{p(z_t)} \left[\prod_{m=1}^{M} \frac{p(z_t | x_t^m) p(x_t^m)}{p(z_t)}\right] p(z_t | z_{t-1}) \\
&\propto p(z_t | x_{t+1:T}^{1:M}) \left[\prod_{m=1}^{M} \frac{p(z_t | x_t^m)}{p(z_t)}\right] \frac{p(z_t | z_{t-1})}{p(z_t)}
\end{aligned}
$$

Future×Present×Past (1)

- Backward Filtering

$$p(z_t|x_{t:T}) \propto p(z_t|x_{t+1:T}) \left[\prod_m \frac{p(z_t|x_t^m)}{p(z_t)} \right]$$

- Forward Smoothing

$$p(z_t|x_{1:T}) \propto p(z_t|x_{t+1:T}) \left[\prod_m \frac{p(z_t|x_t^m)}{p(z_t)} \right] \frac{p(z_t|x_{1:t-1})}{p(z_t)}$$

- Conditional Smoothing Posterior

$$p(z_t|z_{t-1}, x_{t:T}) \propto p(z_t|x_{t+1:T}) \left[\prod_m \frac{p(z_t|x_t^m)}{p(z_t)} \right] \frac{p(z_t|z_{t-1})}{p(z_t)}$$

Future×Present×Past (2)

Each distribution is decomposed into

1. Its dependence on future observations

$$p(z_t | x_{t+1:T})$$

2. Its dependence on each modality m in the present

$$p(z_t | x_t^m)$$

3. Its dependence on the past

$$p(z_t | z_{t-1}) \text{ or } p(z_t | x_{1:t-1})$$

Insights of the Factorizations

- Any missing modalities $\bar{m} \in [1, M]$ at time t can simply be left out of the product over modalities, leaving us with distributions that correctly condition on only the modalities $[1, M] \backslash \{\bar{m}\}$ that are present
- We can compute all three distributions if we can approximate the dependence on the future

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

learn approximate posteriors

$$q(z_t | x_t^m) \simeq p(z_t | x_t^m)$$

for each modality m, and know the model dynamics

$$p(z_t), p(z_t | z_{t-1})$$

Outline

Gaussian Assumption

- It is not tractable to compute the product of generic probability distributions
- So assume that each term in the factorization is Gaussian
- If each distribution is Gaussian, then their products or quotients are also Guassian, and their products or quotients can be computed in closed form

Uncertainty Awareness

- The output distribution of Product-of-Gaussians is dominated by the input Gaussian term with lower variance (higher precision), thereby fusing information in a way that gives more weight to higher-certainty inputs
- Automatically balances the information provided by each modality m, depending on:
  - whether $p(z_t|x_t^m)$ is high or low certainty
  - the information provided from the past and future through $p(z_t|z_{t-1})$ and $p(z_t|x_{t+1:T})$
- Thereby performing multimodal temporal fusion in a manner that is uncertainty-aware

Outline

Missing Observations in the Future

- $p(z_t|x_{t+1:T})$ does not admit further factorization, hence does not readily handle missing data among those future observations
- $z_t \perp\!\!\!\perp x_{t+1:T}|z_{t+1}$ (by d-seperation)

$$
\begin{aligned}
\Rightarrow p(z_t|x_{t+1:T}) &= \int_{z_{t+1}} p(z_t, z_{t+1}|x_{t+1:T})dz_{t+1} \\
&= \int_{z_{t+1}} p(z_t|z_{t+1}, x_{t+1:T})p(z_{t+1}|x_{t+1:T})dz_{t+1} \\
&= \int_{z_{t+1}} p(z_t|z_{t+1})p(z_{t+1}|x_{t+1:T})dz_{t+1} \\
&= \mathbb{E}_{p(z_{t+1}|x_{t+1:T})}[p(z_t|z_{t+1})]
\end{aligned}
$$

Approximating $p(z_t|x_{t+1:T}) = \mathbb{E}_{p(z_{t+1}|x_{t+1:T})}[p(z_t|z_{t+1})]$

- Tractable approximation via Huber et al. 2011
- Assume $p(z_t|x_{t+1:T}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with diagonal $\boldsymbol{\Sigma}$
- Assume $p(z_t|z_{t+1}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with diagonal $\boldsymbol{\Sigma}$
- Draw $(\mu_1, \Sigma_2), \ldots, (\mu_K, \Sigma_K)$ of $p(z_t|z_{t+1})$ under $p(z_{t+1}|x_{t+1:T})$, then
  - Approximate $\hat{\mu}$ of $p(z_t|x_{t+1:T})$ via moment-matching as

$$\frac{1}{K}\sum_{k=1}^{K}\mu_k$$

  - Approximate $\hat{\Sigma}$ of $p(z_t|x_{t+1:T})$ via moment-matching as

$$\frac{1}{K}\sum_{k=1}^{K}(\Sigma_k + \mu_k^2) - \hat{\mu}^2$$

Insights of $p(z_t|x_{t+1:T}) = \mathbb{E}_{p(z_{t+1}|x_{t+1:T})}[p(z_t|z_{t+1})]$ (1)

- The backward filtering distribution

$$p(z_t|x_{t:T}) \propto p(z_t|x_{t+1:T}) \left[\prod_m \frac{p(z_t|x_t^m)}{p(z_t)} \right]$$

becomes

$$p(z_t|x_{t:T}) \propto \mathbb{E}_{p(z_{t+1}|x_{t+1:T})}[p(z_t|z_{t+1})] \left[\prod_{m=1}^{M} \frac{p(z_t|x_t^m)}{p(z_t)} \right]$$

- By sampling under the filtering distribution for time $t+1$, $p(z_{t+1}|x_{t+1:T})$, we can compute the filtering distribution for time t, $p(z_t|x_{t:T})$
- We can recursively compute $p(z_t|x_{t:T})$ backwards in time, starting from $t = T$:

$$p(z_T|x_{T:T}) \rightarrow p(z_{T-1}|x_{T:T}) \rightarrow p(z_{T-1}|x_{T-1:T}) \rightarrow \cdots \rightarrow p(z_1|x_{1:T})$$

Insights of $p(z_t|x_{t+1:T}) = \mathbb{E}_{p(z_{t+1}|x_{t+1:T})}[p(z_t|z_{t+1})]$ (2)

- Once we can perform

$$p(z_t|x_{t:T}) \propto \mathbb{E}_{p(z_{t+1}|x_{t+1:T})}[p(z_t|z_{t+1})] \left[\prod_{m=1}^{M} \frac{p(z_t|x_t^m)}{p(z_t)}\right]$$

filtering backwards in time, we can use this to approximate $p(z_t|x_{t+1:T})$ in the smoothing distribution

$$p(z_t|x_{1:T}) \propto p(z_t|x_{t+1:T}) \left[\prod_{m} \frac{p(z_t|x_t^m)}{p(z_t)}\right] \frac{p(z_t|x_{1:t-1})}{p(z_t)}$$

and the conditional smoothing posterior

$$p(z_t|z_{t-1}, x_{t:T}) \propto p(z_t|x_{t+1:T}) \left[\prod_{m} \frac{p(z_t|x_t^m)}{p(z_t)}\right] \frac{p(z_t|z_{t-1})}{p(z_t)}$$

Insights of $p(z_t|x_{t+1:T}) = \mathbb{E}_{p(z_{t+1}|x_{t+1:T})}[p(z_t|z_{t+1})]$ (3)

- This approach removes the explicit dependence on all future observations $x_{t+1:T}$, allowing us to handle missing data
- Suppose the data points

$$X_{\nexists} = \{x_{t_i}^{m_i}\}$$

are missing, rather than directly compute the dependence on an incomplete set of future observations

$$p(z_t|x_{t+1:T} \backslash X_{\nexists})$$

we can instead sample z_{t+1} under the filtering distribution conditioned on incomplete observations

$$p(z_{t+1}|x_{t+1:T} \backslash X_{\nexists})$$

and then compute $p(z_t|z_{t+1})$ given the sampled z_{t+1}, thereby approximating $p(z_t|x_{t+1:T} \backslash X_{\nexists})$

Outline

Factorized Variational Approximations (1)

- Define the variational posterior approximation q:

$$q(z_t|x_t^m) \equiv \tilde{q}(z_t|x_t^m)p(z_t)$$

- $\tilde{q}(z_t|x_t^m)$ is parameterized by a time-invariant neural network for each modality m
- We learn the Gaussian quotients $\tilde{q}(z_t|x_t^m)$ directly, so as to avoid the constraint required for ensuring a quotient of Gaussians is well-defined:

$$\tilde{q}(z_t|x_t^m) = \frac{q(z_t|x_t^m)}{p(z_t)}$$

- We also parameterize the transition dynamics $p(z_t|z_{t-1})$ and $p(z_t|z_{t+1})$ using neural networks for the quotient distributions

Factorized Variational Approximations (2)

- Denote $\mathbb{E}_{\leftarrow}$ as a shorthand for the expectation under the approximate backward filtering distribution $q(z_{t+1}|x_{t+1:T})$:

$$p(z_t|x_{t+1:T}) = \mathbb{E}_{p(z_{t+1}|x_{t+1:T})}[p(z_t|z_{t+1})] = \mathbb{E}_{\leftarrow}[p(z_t|z_{t+1})]$$

- Denote $\mathbb{E}_{\rightarrow}$ as the expectation under the forward smoothing distribution $q(z_{t-1}|x_{1:T})$:

$$p(z_t|x_{1:t-1}) = \mathbb{E}_{q(z_{t-1}|x_{1:T})}[p(z_t|z_{t-1})] = \mathbb{E}_{\rightarrow}[p(z_t|z_{t-1})]$$

Factorized Variational Approximations (3)

1. Backward Filtering (Variational Backward Algorithm)

$$q(z_t|x_{t:T}) \propto \mathbb{E}_{\leftarrow}[p(z_t|z_{t+1})] \prod_m \tilde{q}(z_t|x_t^m)$$

2. Forward Smoothing (Variational Backward-Forward Algorithm)

$$q(z_t|x_{1:T}) \propto \mathbb{E}_{\leftarrow}[p(z_t|z_{t+1})] \prod_m \tilde{q}(z_t|x_t^m) \frac{\mathbb{E}_{\rightarrow}[p(z_t|z_{t-1})]}{p(z_t)}$$

3. Conditional Smoothing Posterior

$$q(z_t|z_{t-1}, x_{t:T}) \propto \mathbb{E}_{\leftarrow}[p(z_t|z_{t+1})] \prod_m \tilde{q}(z_t|x_t^m) \frac{p(z_t|z_{t-1})}{p(z_t)}$$

Variational Backward Algorithm

function $\textsc{BackwardFilter}(x_{1:T}, K)$
  Initialize $q(z_t|x_{T+1:T}) \leftarrow p(z_T)$
  for $t = T$ to 1 **do**
    Let $\mathcal{M} \subset [1, M]$ be the observed modailities at t
    $q(z_t|x_{t:T}) \leftarrow q(z_t|x_{t+1:T}) \prod_{\mathcal{M}} \tilde{q}(z_t|x_t^m)$
    Sample K particles $z_t^k \sim q(z_t|x_{t:T})$ for $k \in [1, K]$
    Compute $p(z_{t-1}|z_t^k)$ for each particle z_t^k
    $q(z_{t-1}|x_{t:T}) \leftarrow \frac{1}{K} \sum_{k=1}^{K} p(z_{t-1}|z_t^k)$
  end for
  return $\{q(z_t|x_{t:T}), q(z_t|x_{t+1:T})$ for $t \in [1, T]\}$
end function

Variational Backward Algorithm (Remarks)

- By reversing time:
  - The algorithm gives us a variational forward algorithm that computes the forward filtering distribution

$$q(z_t | x_{1:t})$$

- By setting the number of particles $K = 1$:
  - The algorithm effectively computes the conditional filtering posterior

$$q(z_t | z_{t+1}, x_t)$$

    and conditional prior

$$p(z_t | z_{t+1})$$

    for a randomly sampled latent sequence $z_{1:T}$

Variational Backward-Forward Algorithm

function $\textsc{ForwardSmooth}(x_{1:T}, K_b, K_f)$
  Initialize $\tilde{p}(z_t|x_{1:0}) \leftarrow 1$
  Collect $q(z_t|x_{t+1:T})$ from $\textsc{BackwardFilter}(x_{1:T}, K_b)$
  for $t = 1$ to T **do**
    Let $\mathcal{M} \subset [1, M]$ be the observed modalities at t
    $q(z_t|x_{1:T}) \leftarrow q(z_t|x_{t+1:T}) \prod_{\mathcal{M}} [\tilde{q}(z_t|x_t^m)] \frac{q(z_t|x_{1:t-1})}{p(z_t)}$
    Sample K_f particles $z_t \sim q(z_t|x_{1:T})$ for $k \in [1, K_f]$
    Compute $p(z_{t+1}|z_t^k)$ for each particle z_t^k
    $q(z_{t+1}|x_{1:t}) \leftarrow \frac{1}{K_f} \sum_{k=1}^{K_f} p(z_{t+1}|z_t^k)$
  end for
  return $\{q(z_t|x_{1:T}), q(z_t|x_{1:t-1})$ for $t \in [1, T]\}$
end function

Variational Backward-Forward Algorithm (Remarks)

- By setting the number of particles $K_f = 1$:
  - The algorithm effectively computes the conditional smoothing posterior
    $$q(z_t | z_{t-1}, x_{t:T})$$
    and conditional prior
    $$p(z_t | z_{t-1})$$
    for a randomly sampled latent sequence $z_{1:T}$

Knowing $p(z_t)$ of Each t

- Variational Backward-Forward Algorithm requires knowing $p(z_t)$ for each t
- ~~Sampling $p(z_t)$ in the forward pass~~
- We avoid the instability of sampling T successive latents with no observations by instead assuming $p(z_t)$ is constant with time, i.e. the MDMM is stationary when nothing is observed
- During training, we add

$$\mathrm{KL}\left(p(z_t) \| \mathbb{E}_{z_{t-1}} p(z_t|z_{t-1})\right) + \mathrm{KL}\left(p(z_t) \| \mathbb{E}_{z_{t+1}} p(z_t|z_{t+1})\right)$$

to the loss to ensure that the transition dynamics obey this assumption

ELBO for Backward Filtering

- The filtering ELBO:

$$L_{\text{filter}} = \sum_{t=1}^{T} [\mathbb{E}_{q(z_t|x_{t:T})} \log p(x_t|z_t) - \mathbb{E}_{q(z_{t+1}|x_{t+1:T})} \mathsf{KL}(q(z_t|z_{t+1}, x_t) \| p(z_t|z_{t+1}))]$$

- It corresponds to a "backward filtering" variational posterior

$$q(z_{1:T}|x_{1:T}) = \prod_t q(z_t|z_{z+1}, x_t)$$

where each z_t is only inferred using the current observation x_t and the future latent state z_{t+1}

ELBO for Forward Smoothing

- The smoothing ELBO:

$$L_{\text{smooth}} = \sum_{t=1}^{T} [\mathbb{E}_{q(z_t|x_{1:T})} \log p(x_t|z_t) - \mathbb{E}_{q(z_{t-1}|x_{1:T})} \mathrm{KL}(z_t|z_{t-1}, x_{t:T}) \| p(z_t|z_{t-1}))]$$

- It corresponds to the correct factorization of the posterior

$$p(z_{1:T}|x_{1:T}) = p(z_1|x_{1:T}) \prod_{t=2}^{T} p(z_t|z_{t-1}, x_{t:T})$$

where each term combines information from both past and future

Backward-Forward Variational Inference (BFVI)

- Since L_{smooth} corresponds to the correct factorization, it should theoretically be enough to minimize L_{smooth} to learn good MDMM parameters θ, ϕ
- However, in order to compute L_{smooth}, we must perform a backward pass which requires sampling under the backward filtering
- Hence, to accurately approximatee L_{smooth}, the backward filtering distribution has to be reasonably accurate as well
- This motivates learning the parameters θ, ϕ by jointly maximizing the filtering and smoothing ELBOs as a weighted sum
- We call this paradigm BFVI due to its use of variational posteriors for both backward filtering and forward smoothing

Outline

1. Methods
   - Factorized Posterior Distributions
   - Multimodal Fusion via Product of Gaussians
   - Approximate Filtering with Missing Data
   - Backward-Forward Variational Inference
2. Experiments
   - Datasets
   - Inference Tasks
   - Weakly Supervised Learning

MTS Dataset I: Noisy Spirals

- $R \in 2^{\mathcal{U}_{[-1,1)}}$
- $\mathbf{x}(t) : \{0, 1, 2 \dots 99\} \to \mathbb{R}^2$:

$$\mathbf{x}(t) \equiv \begin{bmatrix} \sqrt{R} \cdot r(t)\cos\theta(t) + 0.1 \cdot \mathcal{N} \\ \frac{1}{\sqrt{R}} \cdot r(t)\sin\theta(t) + 0.1 \cdot \mathcal{N} \end{bmatrix}$$

- $r(0) \dots r(99), \theta(0) \dots \theta(99)$:

$$r(0) \equiv 0.25 + \mathcal{U}_{[0,0.5)} \dots r(99) \equiv 2.25 + \mathcal{U}_{[0,0.5)}$$

$$\theta(0) \equiv \mathcal{U}_{[0,\pi)} \dots \theta(99) \equiv \mathcal{U}_{[4\pi,5\pi)}$$

or

$$\theta(0) \equiv \mathcal{U}_{[0,-\pi)} \dots \theta(99) \equiv \mathcal{U}_{[-4\pi,-5\pi)}$$

- 5 latent dimensions
- 2 perceptron layers for encoding $q(z_t|x_t^m)$ and decoding $p(x_t^m|z_t)$

MTS Dataset II: Weizmann Human Actions

- 90 videos of 9 people each performing 10 actions
- We converted it to a trimodal time series dataset by treating silhouette masks and an additional modality, and treating actions as per-frame labels
- We selected one person's videos as the test set, and the other 80 videos as the training set, allowing us to test action label prediction on an unseen person
- 256 latent dimensions
- Convolutional / Deconvolutional neural networks for encoding and decoding

Outline

1. Methods
   - Factorized Posterior Distributions
   - Multimodal Fusion via Product of Gaussians
   - Approximate Filtering with Missing Data
   - Backward-Forward Variational Inference
2. Experiments
   - Datasets
   - Inference Tasks
   - Weakly Supervised Learning

Temporal Inference Tasks

1. Reconstruction: reconstruction given complete observations
2. Drop Half: reconstruction after half of the inputs are randomly deleted
3. Forward Extrapolation: predicting the last 25% of a sequence when the reset is given
4. Backward Extrapolation: inferring the first 25% of a sequence when the reset is given

Weizmann Human Actions

- Multimodal training
- Unimodal testing: we provided only video frames as input
  - NO silhouette masks
  - NO action labels

Cross-Modal Inference Tasks

1. Conditional Generation for Spirals: given x coordinates and initial 25% of y coordinates, generate reset of spirals
2. Conditional Generation for Weizmann: given the video frames, generate the silhouette masks
3. Label Prediction for Weizmann: infer action labels given only video frames

BFVI vs RNN-based Methods

- F-Mask and F-Skip
  - Use forward RNNs, one per modality
  - Use zero-masking and update skipping respectively
- B-Mask and B-Skip
  - Use backward RNNs
  - With masking and skipping respectively
- BFVI achieves high performance on all tasks, whereas RNN-based methods only perform well on a few; in particular, all methods besides BFVI do poorly on the conditional generation task
- RNN lack a principled approach to multimodal fusion, and hence fail to learn a latent space which captures the mutual information between action labels and images
- BFVI learns to both predict one modality from another, and to propagate informatiokn across time

Outline

Two Forms of Weakly Supervised Learning

- Learning with data missing uniformly at random
  - Noisy sensors
  - Asynchronous sensors
- Learning with missing modalities
  - Semi-supervised learning
  - The dataset is partially unlabelled by annotators
  - A fraction of the sequences in the dataset only has a single modality present
  - Sensor break-down