Structured Inference Networks for Nonlinear State Space Models

Rahul G. Krishnan, Uri Shalit, David Sontag

New York University

30 Sep 2016

Chris Cremer

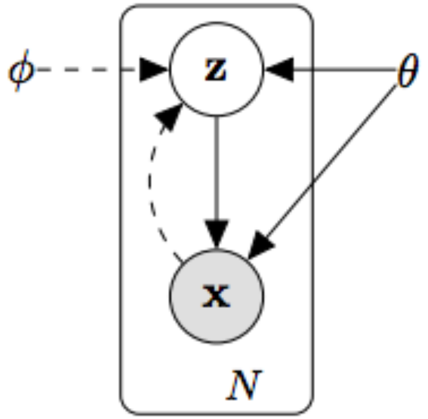
CSC2541

Nov 4 2016

Overview

- VAE
- Gaussian State Space Models
- Inference Network
- Results

Recap - VAE



Generative Model

$$p_\theta(x|z) = \mathcal{N}\big(\mu_\theta(z), \Sigma_\theta(z)\big)$$
$$p_\theta(z) = \mathcal{N}(0, I)$$

Recognition Network

$$q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$$

Use MLP to model the mean and covariance

Learning and Inference –> Maximize Lower Bound

$$\log p_\theta(x) \geq \mathop{\mathbb{E}}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \mathrm{KL}(\, q_\phi(z|x) || p_\theta(z)\,)$$

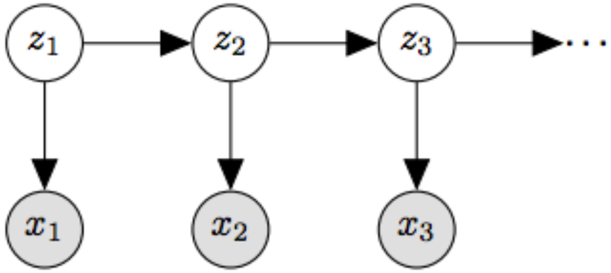
Reconstruction Loss

Calculated by sampling $q_\phi(z|x)$ with reparameterization trick

Divergence from Prior

Analytic equation

Gaussian State Space Models

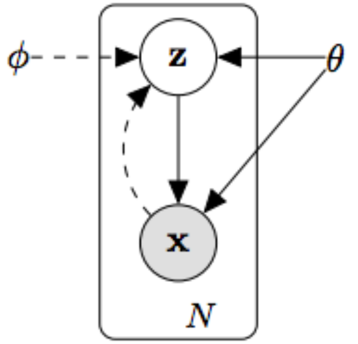


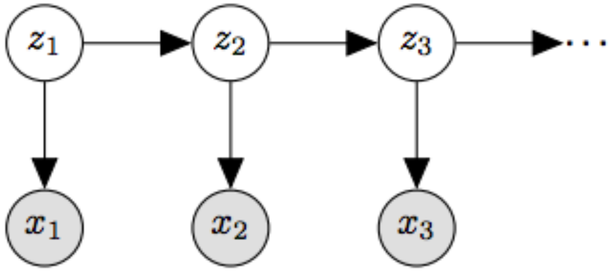
Generative Model

$$z_t \sim \mathcal{N}(G_\alpha(z_{t-1}, \Delta_t), S_\beta(z_{t-1}, \Delta_t)) \textit{ (Transition)} \qquad x_t \sim \Pi(F_\kappa(z_t)) \textit{ (Emission)}$$

- HMM with continuous hidden state
- If transition and emission are linear Gaussian, then we can do inference analytically (Kalman Filter)
- Deep Markov Model:
  - Transition and emissions distributions are parametrized by MLPs
  - Inference: VAE

Inference – Factorized Lower Bound





$$\log p_\theta(x) \geq \mathop{\mathbb{E}}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \mathrm{KL}(\, q_\phi(z|x) || p_\theta(z)\,)$$

Reconstruction Loss

Calculated by sampling $q_\phi(z|x)$ with reparameterization trick

Divergence from Prior

Analytic equation

$$\log p_\theta(\vec{x}) \geq \mathop{\mathbb{E}}_{q_\phi(\vec{z}|\vec{x})} [\log p_\theta(\vec{x}|\vec{z})] - \mathrm{KL}(q_\phi(\vec{z}|\vec{x}) || p_\theta(\vec{z}))$$

$$= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_\phi(z_t|\vec{x})} [\log p_\theta(x_t|z_t)] - \mathrm{KL}(q_\phi(z_1|\vec{x}) || p_\theta(z_1)) - \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_\phi(z_{t-1}|\vec{x})} [\mathrm{KL}(q_\phi(z_t|z_{t-1}, \vec{x}) || p_\theta(z_t|z_{t-1}))]$$

Reconstruction Loss

Calculated by sampling $q_\phi(z_t|\vec{x})$ with reparameterization trick

Divergence from Prior

Analytic equation

Divergence from Prior

Analytic equation

Inference Networks

- Evaluate possibilities for the inference networks
  - Mean-Field Model (MF) vs Structured Model (ST)
  - Observations from past (L), future (R), or both (LR)
- Combiner Function: MLP that combines the previous state with the RNN output

$$\log p_\theta(\vec{x}) \geq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_\phi(z_t|\vec{x})} [\log p_\theta(x_t|z_t)] - \mathrm{KL}(q_\phi(z_1|\vec{x})||p_\theta(z_1)) - \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_\phi(z_{t-1}|\vec{x})} [\mathrm{KL}(q_\phi(z_t|z_{t-1},\vec{x})||p_\theta(z_t|z_{t-1}))]$$

Inf. Network	Variational Approximation	Implementation
MF-LR	$q(z_t\|x_1,\ldots x_T)$	BRNN
MF-L	$q(z_t\|x_1,\ldots x_t)$	RNN
ST-L	$q(z_t\|z_{t-1},x_1,\ldots x_t)$	RNN & comb.fxn
Deep Kalman Smoothing (ST-R) **DKS**	$q(z_t\|z_{t-1},x_t,\ldots x_T)$	RNN & comb.fxn
ST-LR	$q(z_t\|z_{t-1},x_1,\ldots x_T)$	BRNN & comb.fxn

Inference Networks Results

Inf. Network	Variational Approximation	Implementation
MF-LR	$q(z_t \mid x_1, \ldots x_T)$	BRNN
MF-L	$q(z_t \mid x_1, \ldots x_t)$	RNN
ST-L	$q(z_t \mid z_{t-1}, x_1, \ldots x_t)$	RNN & comb.fxn
DKS	$q(z_t \mid z_{t-1}, x_t, \ldots x_T)$	RNN & comb.fxn
ST-LR	$q(z_t \mid z_{t-1}, x_1, \ldots x_T)$	BRNN & comb.fxn

Polyphonic music data (Boulanger-Lewandowski et al., 2012)

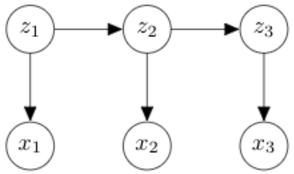
- Sequence of 88-dimensional binary vectors corresponding to the notes of a piano
- Report held-out negative log-likelihood (NLL)

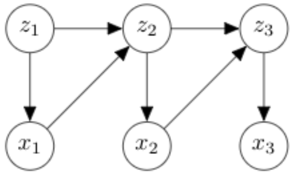
Inference Network	JSB	Nottingham	Piano	Musedata
DKS (i.e., **ST-R**)	6.605 (7.033)	3.136 (3.327)	8.471 (8.584)	7.280 (7.136)
ST-L	7.020 (7.519)	3.446 (3.657)	9.375 (9.498)	8.301 (8.495)
ST-LR	6.632 (7.078)	3.251 (3.449)	8.406 (8.529)	7.127 (7.268)
MF-LR	6.701 (7.101)	3.273 (3.441)	9.188 (9.297)	8.760 (8.877)

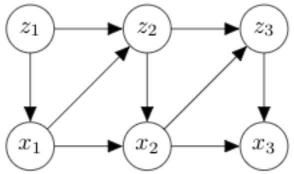
Results:

- **ST-LR** and **DKS** substantially outperform **MF-LR** and **ST-L**
    - Due to previous state (z_{t-1}) and future observations(x_t, …, x_T)
- z_{t-1} summarizes past observations (x_1, …, x_t)
- **DKS** network has half the parameters of the **ST-LR**

Model Comparison







DMM (DKS)

STORN

DMM-Aug (DKS)

TSBN

HMSBN

LV-RNN (NASMC)

Held-out negative log-likelihood (NLL)

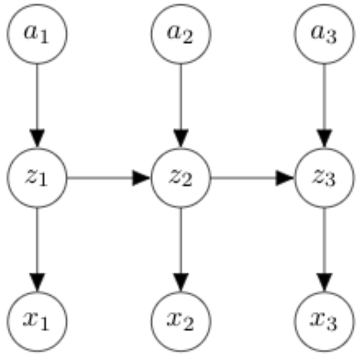
Methods	JSB	Nottingham	Piano	Musedata
DMM	6.388 (6.926) {6.856}	2.770 (2.964) {2.954}	7.835 (7.980) {8.246}	6.831 (6.989) {6.203}
DMM-Aug.	6.288 (6.773) {6.692}	2.679 (2.856) {2.872}	7.591 (7.721) {8.025}	6.356 (6.476) {5.766}
HMSBN	(8.0473) {7.9970}	(5.2354) {5.1231}	(9.563) {9.786}	(9.741) {8.9012}
STORN	6.91	2.85	7.13	6.16
RNN	8.71	4.46	8.37	8.13
TSBN	{7.48}	{3.67}	{7.98}	{6.81}
LV-RNN	3.99	2.72	7.61	6.89

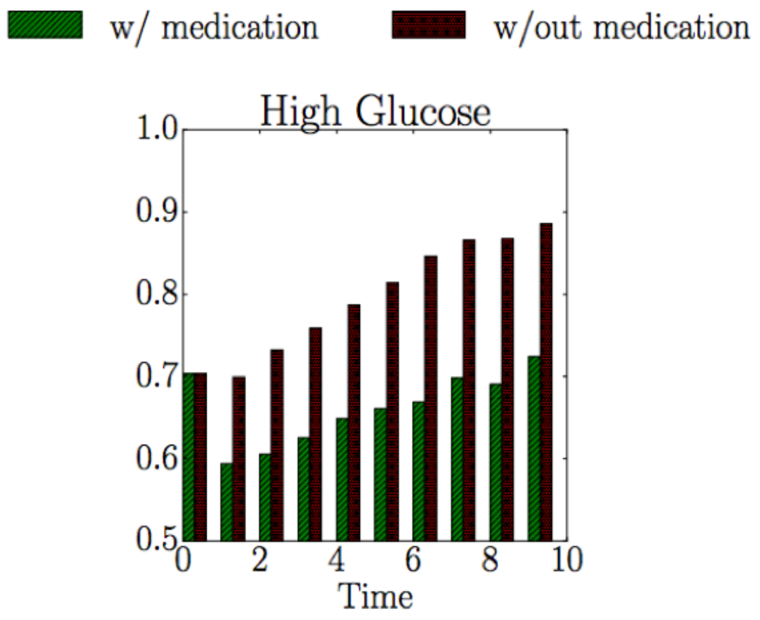
Results:

- Increasing the complexity of the generative model improves the likelihood (DMM vs DMM-Aug)
- DMM-Aug (DKS) obtains better results on all datasets (except LV-RNN on JSB)
- Demonstrates the inference network’s ability to learn powerful generative models

EHR Patient Data

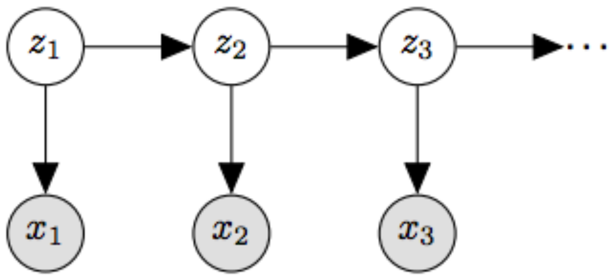
- What would happen if the patient received diabetic medication or not?





Conclusion

- Structured Inference Networks for Nonlinear State Space Models



$$\mathcal{L}(\vec{x};(\theta,\phi)) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_\phi(z_t|\vec{x})} \left[\log p_\theta(x_t|z_t)\right] - \mathrm{KL}(q_\phi(z_1|\vec{x})||p_\theta(z_1))$$
$$- \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_\phi(z_{t-1}|\vec{x})} \left[\mathrm{KL}(q_\phi(z_t|z_{t-1},\vec{x})||p_\theta(z_t|z_{t-1}))\right].$$

VAE for sequential data

Questions?