Hierarchical Deep Generative Models for Multi-Rate Multivariate Time Series

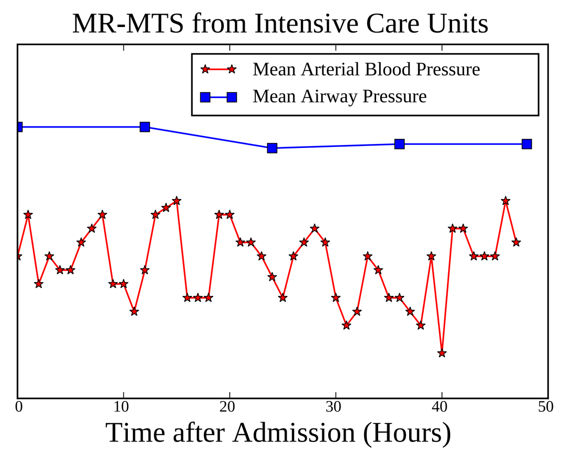
Zhengping Che*, Sanjay Purushotham*, Guangyu Li*, Bo Jiang, Yan Liu

Department of Computer Science, University of Southern California *(*equal contributions)*

{zche, spurusho, guangyul, boj, yanliu.cs}@usc.edu

Background & Motivation

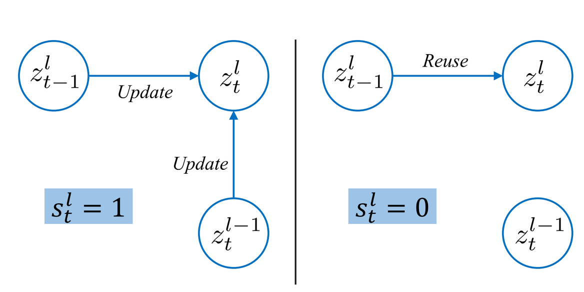
- MR-MTS (multi-rate multivariate time series)
  - Time series with different sampling rates
  - May from multiple data sources/sensors
  - Many real-world applications
    - ∗ Healthcare: vital signs in ICU/lab tests
    - ∗ Climate: daily/seasonal observations
    - ∗ Financial forecasting, mechanical maintenance, business analysis...



- Modeling MR-MTS is *challenging*
  - Data of different sampling rates
  - Multi-scale temporal dependencies
  - Complex underlying generation mechanism
- How can we effectively forecast/interpolate unobserved values in MR-MTS? (E.g., $\boldsymbol{x}_{1:T}^{1:L}$: MR-MTS observations of L sampling rates and T time steps)
  - *Single rate* models? *(Kalman Filter, VAR, DMM, etc.)*
    - ∗ Ignoring dependencies across different rates
  - Simple *imputations*? *(MICE, MissForest, etc.)*
    - ∗ May introduce unrelated or hide natural dependencies
  - Multi-rate *discriminative* models? *(PLSTM, HM-RNN, etc.)*
    - ∗ Not able to learn how the data is generated
- **Our solution**
  - **MR-HDMM: hierarchical deep generative models for MR-MTS!**

Model at a Glance

- **MR-HDMM** – **M**ulti-**R**ate **H**ierarchical **D**eep **M**arkov **M**odel
  - Capturing underlying data generation process
  - Learned by variational inference methods
- Key components to learn the *latent hierarchical structures* of MR-MTS
  - **Learnable switches**
    - ∗ Goal: to let higher-layer states act as summarized representations
    - ∗ **Solution**: an *update-and-reuse* mechanism
    - ∗ Switches will trigger updates only if enough information is got from lower layers



  - **Auxiliary connections**
    - ∗ Goal: to effectively capture multi-scale temporal dependencies in MR-MTS
    - ∗ **Solution**: connecting higher latent layers to lower rate time series
    - ∗ Multi-scale dependencies in lower-rate MTS will not be masked by higher-rate MTS through bottom-up connections in the model
- Jointly learning all parameters by *stochastic backpropagation*[1] and *ancestral sampling*

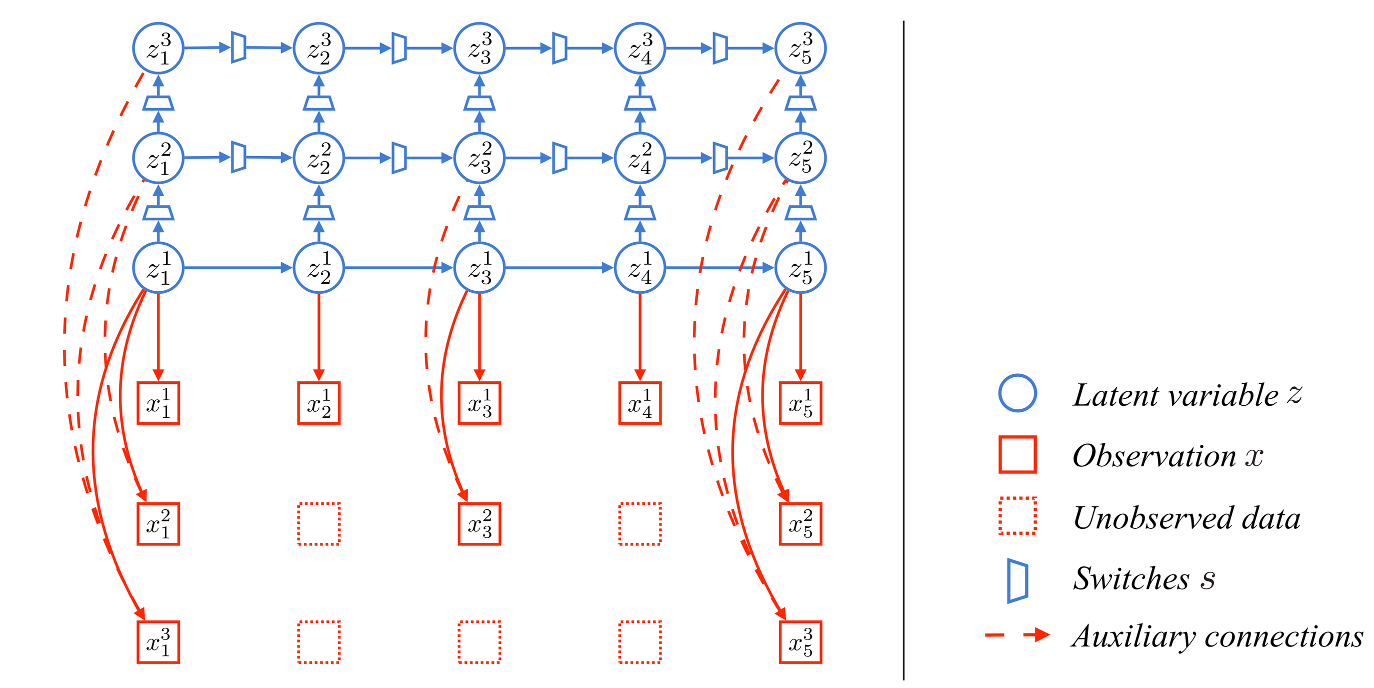
References & Acknowledgments

1. Kingma, Diederik P., et al. *Auto-encoding variational bayes*. ICLR 2014.
2. Johnson, Alistair EW, et al. *MIMIC-III, a freely accessible critical care database*. Sci. Data 2016.
3. Menne, M. J., et al. *Long-term daily and monthly climate records from stations across the contiguous United States*. CDIAC 2010.
4. Krishnan, Rahul G., et al. *Deep kalman filters*. arXiv:1511.05121.
5. Chung, Junyoung, et al. *Hierarchical multiscale recurrent neural networks*. arXiv:1609.01704.
6. Neil, Daniel, et al. *Phased LSTM: Accelerating recurrent network training for long or event-based sequences*. NIPS 2016.

This work is supported in part by NSF Research Grant IIS-1254206 and IIS-1539608, and MURI Grant W911NF-11-1-0332. The views and conclusions should not be interpreted as representing the official policies of the funding agency or the U.S. Government.

Model Part I: Generation Model

- A generation model (θ) with **transition** and **emission** framework



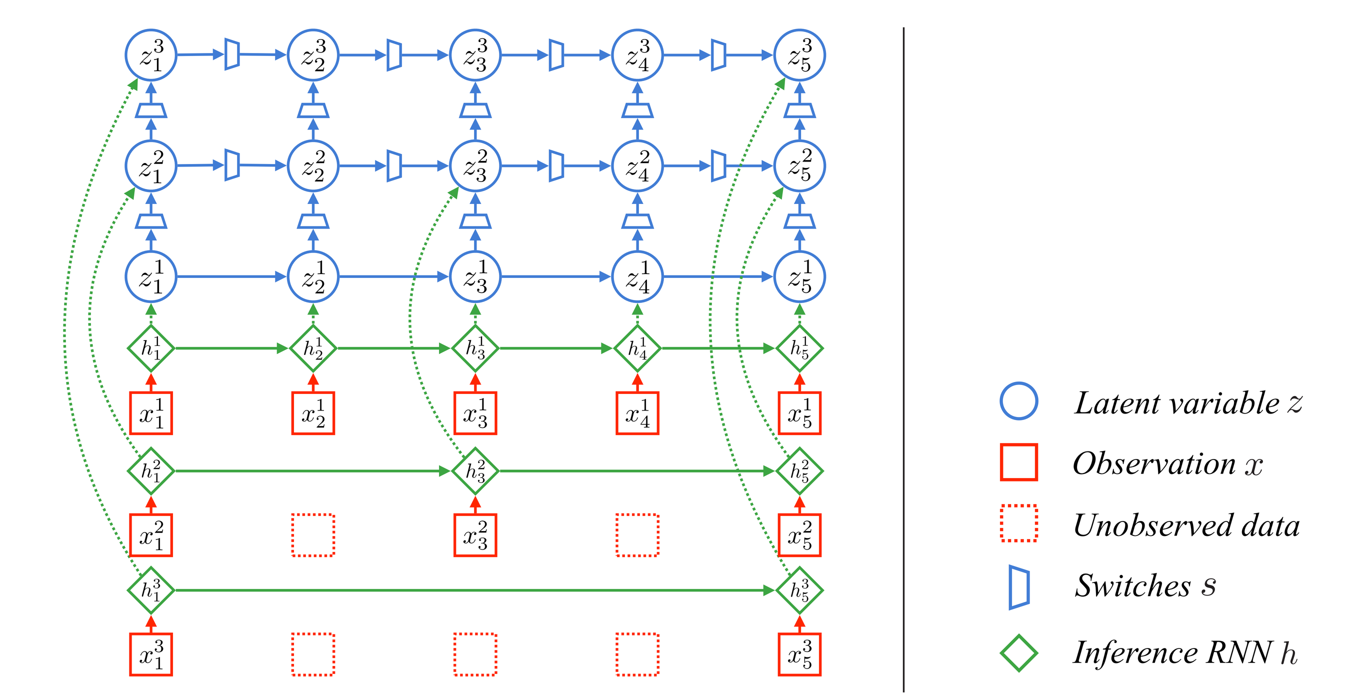
- **Transition** – using states $\boldsymbol{z}$ to capture the latent temporal dependencies
  - Transition distribution of $\boldsymbol{z}$: multivariate Gaussian
- **Emission** – generating observations $\boldsymbol{x}$ from states $\boldsymbol{z}$
  - Emission distribution of $\boldsymbol{x}$: multinomial/Gaussian for discrete/continuous data
- Joint probability of MR-MTS observations and latent states/switches

$$p_\theta\left(\boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{z}_0^{1:L}\right) = \prod_{t=1}^T \prod_{l=1}^L p_{\theta_x}\left(\boldsymbol{x}_t^l | \boldsymbol{z}_t^{1:l}\right) \cdot \prod_{t=1}^T p_{\theta_z}\left(\boldsymbol{z}_t^1 | \boldsymbol{z}_{t-1}^1\right) \cdot \prod_{t=1}^T \prod_{l=2}^L p_{\theta_s}\left(s_t^l | \boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^{l-1}\right) p_{\theta_z}\left(\boldsymbol{z}_t^l | \boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^{l-1}, s_t^l\right)$$

  - Solving marginal MLE? ⇒ Stochastic variational inference!

Model Part II: Inference Network

- An inference network (ϕ) to mimic the structure of the generation model



- Goal: to maximize the variational *evidence lower bound* (ELBO)

$$\mathbb{E}_{q_\phi}\left[\log p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_{0:T}^{1:L}\right)\right] - D_{\mathrm{KL}}\left(q_\phi\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right) \middle\| p_\theta\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{z}_0^{1:L}\right)\right)$$

- Network design for structured and powerful approximations to the posterior
  - Keeping the Markov properties and the same distribution type of z as θ in ϕ
  - Inheriting switches s from θ to ϕ ($\phi_s = \theta_s$)
  - Capturing MR-MTS observations by using multiple RNNs

Network Type	Usage	Input for h_t^l	Variational Approximation for z_t^l
Forward RNN *(filtering)*	Forecasting	$\boldsymbol{x}_{1:t}^l$	$q_\phi\left(\boldsymbol{z}_t^l \| \boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^{l-1}, s_t^l, \boldsymbol{x}_{1:t}^{1:L}\right)$
Bi-directional RNN	Interpolation	$\boldsymbol{x}_{1:T}^l$	$q_\phi\left(\boldsymbol{z}_t^l \| \boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^{l-1}, s_t^l, \boldsymbol{x}_{1:T}^{1:L}\right)$

- The final factorized function to optimize: a summation of expectations of
  - Conditional loglikelihood $\sum_{t=1}^T \sum_{l=1}^L \mathbb{E}_{Q^*(\boldsymbol{z}_t^{1:l})} \log p_{\theta_x}\left(\boldsymbol{x}_t^l | \boldsymbol{z}_t^{1:l}\right)$
  - KL terms over time steps t and layers l

$$\sum_{t=1}^T \mathbb{E}_{Q^*(\boldsymbol{z}_{t-1}^1)} D_{\mathrm{KL}}\left(q_\phi\left(\boldsymbol{z}_t^1 | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{t-1}^1\right) \middle\| p_\theta\left(\boldsymbol{z}_t^1 | \boldsymbol{z}_{t-1}^1\right)\right) + \sum_{t=1}^T \sum_{l=2}^L \mathbb{E}_{Q^*(\boldsymbol{z}_{t-1}^1, \boldsymbol{z}_t^{l-1})} D_{\mathrm{KL}}\left(q_\phi\left(\boldsymbol{z}_t^l | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^{l-1}\right) \middle\| p_\theta\left(\boldsymbol{z}_t^l | \boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^{l-1}\right)\right)$$

Quantitative Results

- Two real-world datasets from healthcare and climate domains

Dataset Name	# of Samples	Sampling Rates (HSR/MSR/LSR)	# of Variables of Each Rate	Time Series Length
MIMIC-III[2]	10 709	1/4/12 Hours	7/11/44	72 Hours
USHCN[3]	100	1/5/10 Days	70/69/69	365 Days

- Forecasting performance on USHCN *(Mean Squared Error(MSE))*

	Method \ Rate	All	HSR	MSR	LSR
Single-Rate Baselines	**Kalman Filter** (KF)	1.236	1.254	1.190	1.148
	Vector Autoregression (VAR)	2.415	2.579	1.921	1.748
	Deep Markov Model (DMM)[4]	0.795	0.608	0.903	0.877
	HM-RNN[5]	0.692	0.594	1.151	**0.775**
	LSTM	0.849	0.688	0.934	0.928
	PLSTM[6]	0.813	0.710	0.870	0.915
Multi-Rate Baselines	**Multiple KF**	1.212	1.082	1.727	1.518
	Multi-Rate KF	0.628	0.542	0.986	0.799
	Multi-Rate DMM (MR-DMM)	0.667	0.611	0.847	0.875
	Hierarchical DMM (HDMM)	0.626	0.568	0.815	0.836
	MR-HDMM	**0.591**	**0.541**	**0.742**	0.795

- Interpolation performance *(Mean Squared Error(MSE))*

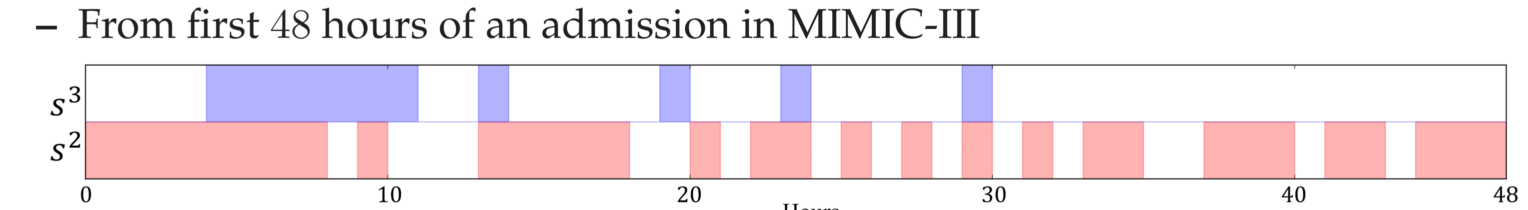
	Method \ Dataset	MIMIC-III In-Sample	MIMIC-III Out-Sample	USHCN In-Sample
Imputation Baselines	**Simple-Mean**	3.812	3.123	0.987
	CubicSpline	3.713	3.212×10^4	0.947
	MICE	3.747	7.580×10^2	0.670
	MissForest	3.863	3.027	0.941
	SoftImpute	3.715	3.086	0.759
Deep Learning Baselines	**DMM**	3.714	3.027	0.782
	MR-DMM	3.710	3.021	0.696
	HDMM	3.790	3.100	0.750
	MR-HDMM	**3.582**	**2.921**	**0.626**

- Lower bound of log-likelihood for all generative models (higher values are better)

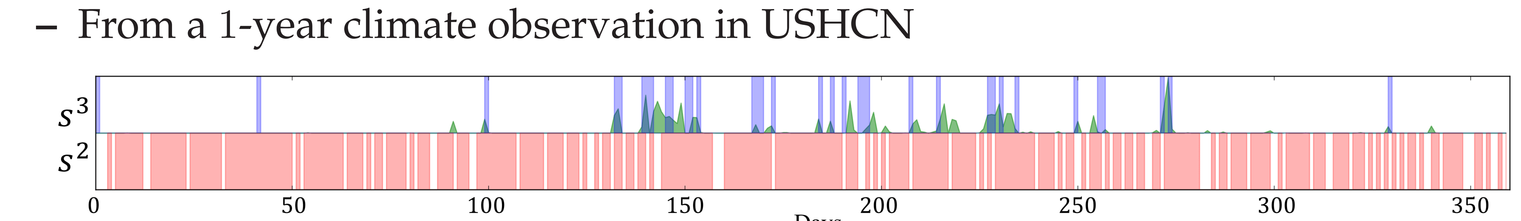
	DMM	MR-DMM	HDMM	MR-HDMM
MIMIC-III	−1.54	2.62	10.54	**15.27**
USHCN	2.37	14.37	17.25	**33.62**

Visualizations

- Latent hierarchical structure learned by MR-HDMM
  - From first 48 hours of an admission in MIMIC-III



  - From a 1-year climate observation in USHCN



  - Blue and Red part: *update*; White part: *reuse*
    - ∗ Higher layers update less frequently and capture longer-term dependencies
  - Green histograms: precipitation time series
    - ∗ Precipitations ⇒ Significant temporal changes captured by the higher layer