Supplementary Materials for Hierarchical Deep Generative Models for Multi-Rate Multivariate Time Series

Zhengping Che [*1] **Sanjay Purushotham** [*1] **Guangyu Li** [*1] **Bo Jiang** [1] **Yan Liu** [1]

A. Derivations

A.1. Getting the ELBO in Equation (2)

To utilize the variational principle and get the ELBO, we introduce another distribution $\mathcal{Q} = q_\phi\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right)$ to approximate $\mathcal{P} = p_\theta\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right)$. Starting with the KL-divergence from $\mathcal{Q}$ to $\mathcal{P}$, we have

$$
\begin{aligned}
0 \leq & D_{\mathrm{KL}}\left(\mathcal{Q} \middle\| \mathcal{P}\right) \\
= & \mathbb{E}_{\mathcal{Q}} \log \frac{\mathcal{Q} \cdot p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_0^{1:L}\right)}{p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L}, \boldsymbol{z}_0^{1:L}\right) \cdot p_\theta\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{z}_0^{1:L}\right)} \\
= & \mathbb{E}_{\mathcal{Q}}\left[\log \frac{\mathcal{Q}}{p_\theta\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{z}_0^{1:L}\right)} - \log p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_{0:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L}\right)\right] \\
& + \log p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_0^{1:L}\right) \\
= & D_{\mathrm{KL}}\left(\mathcal{Q} \middle\| p_\theta\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{z}_0^{1:L}\right)\right) \\
& - \mathbb{E}_{\mathcal{Q}}\left[\log p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_{0:T}^{1:L}\right)\right] + \log p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_0^{1:L}\right)
\end{aligned}
$$

There are two things to be noticed about this equation. First, the expectations are under $\mathcal{Q} = q_\phi\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right)$ and $p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_0^{1:L}\right)$ does not depend on it. Second, in our generation model $\boldsymbol{x}$ is independent from $\boldsymbol{s}$ given $\boldsymbol{z}$ by its design. Then we have

$$
\begin{aligned}
& \mathcal{L}(\theta) \\
= & \log p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_0^{1:L}\right) \\
\geq & \mathbb{E}_{\mathcal{Q}}\left[\log p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_{0:T}^{1:L}\right)\right] - D_{\mathrm{KL}}\left(\mathcal{Q} \middle\| p_\theta\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{z}_0^{1:L}\right)\right) \\
= & \mathcal{F}(\theta, \phi)
\end{aligned}
$$

We find that the tightness of this bound depends on how well $\mathcal{Q}$ approximates the true posterior of latent variables given the inputs and initial parameters $\mathcal{P}$, because the equality holds if and only if $D_{\mathrm{KL}}\left(\mathcal{Q} \middle\| \mathcal{P}\right) = 0$. This requirement is carefully taken into consideration in our inference model design.

A.2. Factorizing the ELBO in Equation (5)

The first part of the $\mathcal{F}(\theta, \phi)$ can be factorized similarly as Equation (1), as follows

$$
\begin{aligned}
& \mathbb{E}_{\mathcal{Q}} \log p_\theta\left(\boldsymbol{x}_{1:T}^{1:L} | \boldsymbol{z}_{0:T}^{1:L}\right) \\
= & \mathbb{E}_{\mathcal{Q}} \sum_{t=1}^{T} \sum_{l=1}^{L} \log p_{\theta_x}\left(\boldsymbol{x}_t^l | \boldsymbol{z}_t^{1:l}\right) \\
= & \mathbb{E}_{\mathcal{Q}} \sum_{t=1}^{T} \sum_{l=1}^{L} \log p_{\theta_x}\left(\boldsymbol{x}_t^l | \boldsymbol{z}_t^{1:l}\right) \\
= & \sum_{t=1}^{T} \sum_{l=1}^{L} \mathbb{E}_{\mathcal{Q}^*\left(\boldsymbol{z}_t^{1:l}\right)} \log p_{\theta_x}\left(\boldsymbol{x}_t^l | \boldsymbol{z}_t^{1:l}\right)
\end{aligned} \tag{6}
$$

Here, $\mathcal{Q}^*\left(\boldsymbol{z}_t^{1:l}\right)$ is the marginal distribution of $\boldsymbol{z}_t^{1:l}$ in the variational approximation to the posterior $q_\phi\left(\boldsymbol{z}_{1:t}^{1:l} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right)$ and is defined as

$$
\begin{aligned}
& \mathcal{Q}^*\left(\boldsymbol{z}_t^{1:l}\right) \\
= & \int q_\phi\left(\boldsymbol{z}_{1:t}^{1:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right) \mathrm{d}\boldsymbol{z}_{1:t-1}^{1:L} \\
= & \int q_\phi\left(\boldsymbol{z}_t^{1:L} | \boldsymbol{x}_{1:t}^{1:L}, \boldsymbol{z}_{t-1}^{1:L}\right) q_\phi\left(\boldsymbol{z}_{1:t-1}^{1:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right) \mathrm{d}\boldsymbol{z}_{1:t-1}^{1:L} \\
= & \mathbb{E}_{\mathcal{Q}^*\left(\boldsymbol{z}_{t-1}^{1:l}\right)} q_\phi\left(\boldsymbol{z}_t^{1:L} | \boldsymbol{x}_{1:t}^{1:L}, \boldsymbol{z}_{t-1}^{1:L}\right)
\end{aligned}
$$

where $\mathcal{Q}^*\left(\boldsymbol{z}_{t-1}^{1:l}\right)$ denotes the marginal distribution of $\boldsymbol{z}_{t-1}^{1:l}$ in the same way.

The second part of the $\mathcal{F}(\theta, \phi)$ can be factorized based on the principle of minimum discrimination information (MDI). First, we have

$$
\begin{aligned}
& D_{\mathrm{KL}}\left(q_\phi\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right) \middle\| p_\theta\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{z}_0^{1:L}\right)\right) \\
= & \mathbb{E}_{q_\phi} D_{\mathrm{KL}}\left(q_\phi\left(\boldsymbol{z}_T^{1:L}, \boldsymbol{s}_T^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{T-1}^{1:L}\right) \middle\| p_\theta\left(\boldsymbol{z}_T^{1:L}, \boldsymbol{s}_T^{2:L} | \boldsymbol{z}_{T-1}^{1:L}\right)\right) \\
& + D_{\mathrm{KL}}\left(q_\phi\left(\boldsymbol{z}_{1:T-1}^{1:L}, \boldsymbol{s}_{1:T-1}^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right) \middle\| p_\theta\left(\boldsymbol{z}_{1:T-1}^{1:L}, \boldsymbol{s}_{1:T-1}^{2:L} | \boldsymbol{z}_0^{1:L}\right)\right)
\end{aligned} \tag{7}
$$

[*] Equal contribution [1] Department of Computer Science, University of Southern California, Los Angeles, California, United States. Correspondence to: Zhengping Che, Sanjay Purushotham, Guangyu Li, Bo Jiang, Yan Liu <{zche,spurusho,guangyul,boj,yanliu.cs}@usc.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

where q_ϕ is a shorthand for $q_\phi\left(\boldsymbol{z}_{1:T-1}^{1:L}, \boldsymbol{s}_{1:T-1}^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right)$. Notice that $q_{\phi_s}(.)$ and $p_{\theta_s}(.)$ are the same and $\boldsymbol{s}$ is deterministic given other variables and thus can be integrated out (by using its value). Then, the first KL divergence term can be further factorized as

$$
\begin{aligned}
&D_{\mathrm{KL}}\Big(q_\phi\left(\boldsymbol{z}_T^{1:L}, \boldsymbol{s}_T^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{T-1}^{1:L}\right) \Big\| p_\theta\left(\boldsymbol{z}_T^{1:L}, \boldsymbol{s}_T^{2:L} | \boldsymbol{z}_{T-1}^{1:L}\right)\Big) \\
=&D_{\mathrm{KL}}\Big(q_\phi\left(\boldsymbol{z}_T^{1} | \boldsymbol{x}_{T}^{1}, \boldsymbol{z}_{T-1}^{1}\right) \Big\| p_\theta\left(\boldsymbol{z}_T^{1} | \boldsymbol{z}_{T-1}^{1}\right)\Big) \\
&+\sum_{l=2}^{L} \mathbb{E}_{q_\phi} D_{\mathrm{KL}}\Big(q_\phi\left(\boldsymbol{z}_T^{l} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{T-1}^{1}, \boldsymbol{z}_T^{l-1}, s_T^l\right) \Big\| p_\theta\left(\boldsymbol{z}_T^{1} | \boldsymbol{z}_{T-1}^{1}, \boldsymbol{z}_T^{l-1}, s_T^l\right)\Big) \\
&+\sum_{l=2}^{L} \mathbb{E}_{q_\phi} D_{\mathrm{KL}}\Big(q_\phi\left(s_T^{l} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{T-1}^{1}, \boldsymbol{z}_T^{l-1}\right) \Big\| p_\theta\left(s_T^{l} | \boldsymbol{z}_{T-1}^{1}, \boldsymbol{z}_T^{l-1}\right)\Big) \\
=&D_{\mathrm{KL}}\Big(q_\phi\left(\boldsymbol{z}_T^{1} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{T-1}^{1}\right) \Big\| p_\theta\left(\boldsymbol{z}_T^{1} | \boldsymbol{z}_{T-1}^{1}\right)\Big) \\
&+\sum_{l=2}^{L} \mathbb{E}_{q_\phi} D_{\mathrm{KL}}\Big(q_\phi\left(\boldsymbol{z}_T^{l} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{T-1}^{1}, \boldsymbol{z}_T^{l-1}\right) \Big\| p_\theta\left(\boldsymbol{z}_T^{1} | \boldsymbol{z}_{T-1}^{1}, \boldsymbol{z}_T^{l-1}\right)\Big)
\end{aligned}
$$

By recursively factorizing the last term of Equation (7), we have

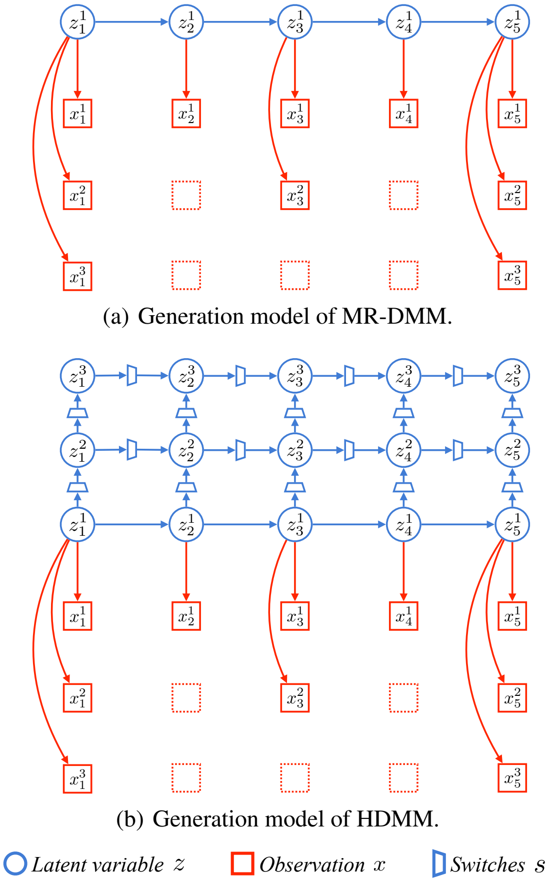
$$
\begin{aligned}
&D_{\mathrm{KL}}\Big(q_\phi\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_0^{1:L}\right) \Big\| p_\theta\left(\boldsymbol{z}_{1:T}^{1:L}, \boldsymbol{s}_{1:T}^{2:L} | \boldsymbol{z}_0^{1:L}\right)\Big) \\
=&\sum_{t=1}^{T} \mathbb{E}_{\mathcal{Q}^*(\boldsymbol{z}_{t-1}^1)} D_{\mathrm{KL}}\Big(q_\phi\left(\boldsymbol{z}_t^{1} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{t-1}^{1}\right) \Big\| p_\theta\left(\boldsymbol{z}_t^{1} | \boldsymbol{z}_{t-1}^{1}\right)\Big) \\
&+\sum_{t=1}^{T}\sum_{l=2}^{L} \mathbb{E}_{\mathcal{Q}^*(\boldsymbol{z}_{t-1}^1, \boldsymbol{z}_t^{l-1})} \\
&\quad D_{\mathrm{KL}}\Big(q_\phi\left(\boldsymbol{z}_t^{l} | \boldsymbol{x}_{1:T}^{1:L}, \boldsymbol{z}_{t-1}^{l}, \boldsymbol{z}_t^{l-1}\right) \Big\| p_\theta\left(\boldsymbol{z}_t^{1} | \boldsymbol{z}_{t-1}^{1}, \boldsymbol{z}_t^{l-1}\right)\Big)
\end{aligned} \tag{8}
$$

where $\mathcal{Q}^*$ is the marginalized distributions defined as previous. Finally, taking both Equation (6) and (8) leads to the factorized ELBO in Equation (5).

B. Baseline Descriptions

B.1. Simplified Models from MR-HDMM

To demonstrate the advantage of the *learnable hierarchical structure* and the *auxiliary connections*, we compared two simplified models derived from the proposed MR-HDMM. The first baseline is named as Multi-Rate Deep Markov Model (MR-DMM), which removes the hierarchical structure in the latent space from the model. The second baseline, which is named as Hierarchical Deep Markov Model (HDMM), removes the auxiliary connections between the lower-rate time series and the higher level latent layers. The other parts of the two models remains the same as MR-HDMM. The generation models of MR-DMM and HDMM are shown in Figure 4(a) and 4(b), respectively.



(a) Generation model of MR-DMM.

(b) Generation model of HDMM.

Figure 4. Generation models of two baseline models derived from the proposed MR-HDMM.

B.2. Implementation Details of KF-based Models

Kalman Filters (KF) We first up-sample all the MSR and LSR features to make their sampling rate the same as the HSR features. We then get single-rate multivariate time series (SR-MTS) for both MIMIC-III dataset and USHCN climate dataset. Then we train KF on the SR-MTS data using EM algorithm to get the forecasting results.

Multiple Kalman Filters (MKF) We train three different KF models on the HSR/MSR/LSR time series separately, and then we concatenate the outputs (eg. forecasting results) of these three KF models to obtain the final results.

Multi-rate Kalman Filters (MR-KF) Three different KF models are trained on the HSR/MSR/LSR time series separately to get their state estimations. Then a neural network (multi-layer perceptron(MLP)) is employed to fuse the estimated state vectors from each Kalman filter to obtain the final prediction results. Similar to the other deep learning methods, MR-KF is trained on the training set, the best weights are chosen based on the performance on the validation set, and the results are reported on the held-out test set.