3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–18

Scalable Hybrid Hidden Markov Model with Gaussian Process Emission for Sequential Time-series Observations

Yohan Jung and **Jinkyoo Park** {BECRE1776, JINKYOO.PARK}@KAIST.AC.KR
Industrial & Systems Engineering, KAIST, Daejeon, Republic of Korea

1. Introduction

Employing non-linear function with Gaussian Process (GP) prior as the emission function of HMM has some advantages in characterizing the sequences of time-series observations; it can model the pair of input-output observations such as observed time and the corresponding response (Frigola, 2015; Nakamura et al., 2017; Nagano et al., 2018). Moreover, employing SM kernel (Wilson and Adams, 2013) for the covariance of GP prior can characterize the dataset based on diversely characterized stationary kernel (Ulrich et al., 2014, 2015). However, due to the scalability issue for the training, this model is limited to modeling a small size of dataset.

In this work, we introduce a hybrid Bayesian HMM with GP emission using SM kernel, which we call HMM-GPSM, for modeling sequences of single-channel time-series observations. Then, we propose a scalable inference method to train the HMM-GPSM with large-scale sequences of time-series data $D = \{x_t, y_t\}_{t=1}^{T}$ with x_t and $y_t \in R^{N_t}$ having (1) long sequences of state transitions (T is large) and (2) a large number of time-series observation from each state (N_t is large).

- To address issue (1), we employ stochastic variational inference (SVI) based on (Hoffman et al., 2013; Foti et al., 2014) to efficiently update the parameters of HMM-GPSM with the long sequences dataset. To be specific, we propose the approximate evidence lower bound for the full T sequences of time-series observations which can be computed using the randomly sampled L sub-sequences of time-series observations. This approximation can linearly reduce the computational complexity for computing the evidence lower bound from $\mathcal{O}(T)$ to $\mathcal{O}(L)$.

- To address issue (2), we propose the approximate GP emission using spectral points sampled from the spectral density of SM kernel and the efficient inference for the kernel hyperparameters of approximate GP emission and corresponding HMM-GPSM. To be specific, we approximate the SM kernel by employing the spectral points sampled from Gaussian mixture spectral density based on random Fourier feature (RFF) (Rahimi and Recht, 2008). We then introduce the variational distribution on the spectral points while treating these points are random variables and derive the regularized lower bound of GP emission likelihood. Using this lower bound, we derive the evidence lower bound of the HMM-GPSM that can be scalably computed for a large number of time-series observations in each sequence. This approximation reduces the computational complexity for computing GP emission likelihood from $\mathcal{O}(N_t^3)$ to $\mathcal{O}(N_t M^2)$ with M sampled spectral points under $M \ll N_t$.

The proposed methods can be jointly used to efficiently update the parameters of HMM-GPSM with the dataset having both (1) and (2) issues. We validate the proposed method on the synthetic using the clustering accuracy and training time as the performance metrics.

2. Methodology

2.1. Hybrid Bayesian HMM with GP Emission

Hybrid Bayesian HMM with GP emission has the structure of Input-Output HMM (Bengio and Frasconi, 1995) with the emission function being modeled by GP. To introduce the model, we assume that T sequences of inputs and outputs pairs $X = \{x_t\}_{t=1}^T$ and $Y = \{y_t\}_{t=1}^T$ are given where $y_t \in R^{N_t}$ is the N_t dimensional output corresponding to the input $x_t \in R^{N_t}$ obtained at time t. We denote $z_t \in \{1, .., K\}$ representing the hidden state for x_t and y_t.

To explain the relation between y_t and x_t, the target function f_t with the conditional GP prior given hidden state z_t and covariance k_{z_t} is defined as

$$p(f_t|x_t, z_t) = N\left(f_t; m_{z_t}(x_t), k_{z_t}(x_t, x_t; \theta_{z_t})\right), \tag{1}$$

where θ_{z_t} denotes the hyperparameters of the SM kernel k_{z_t} having Q_{z_t} mixture components. Under the Gaussian noise assumption, we define the emission function as the conditional marginal likelihood $p(y_t|x_t, z_t)$ given the hidden state z_t and the input x_t as

$$p(y_t|x_t, z_t) = N\left(y_t|m_{z_t}(x_t), k_{z_t}(x_t, x_t; \theta_{z_t}) + \sigma^2_{\epsilon_{z_t}} I\right). \tag{2}$$

Then, the joint likelihood of Hybrid HMM with GP emission is defined as

$$p(Z, Y, A, \pi|X) = p(\pi)p(A)p(z_0|\pi)\prod_{t=1}^{T} p(z_t|z_{t-1}, A)p(y_t|x_t, z_t), \tag{3}$$

where $p(\pi) = \mathrm{Dir}(\pi|\alpha^\pi)$ is a prior distribution for the initial parameter $\pi \in R^K$ and $p(A) = \prod_{i=1}^K \mathrm{Dir}(A_i|\alpha_i^A)$ is a prior distribution for transition matrix $A \in R^{K\times K}$. A_i denotes the i th row of A.

2.2. Variational Inference

Under Mean-Field assumption, let us assume the variational distribution $q(\pi, A, Z) = q(\pi)q(A)q(Z)$,

$$q(\pi) = \mathrm{Dir}(\pi|w^\pi) \qquad q(A) = \prod_{k=1}^{K} \mathrm{Dir}(A_k|w_k^A), \tag{4}$$

where w^π and $\{w_j^A\}_{j=1}^K$ are variational parameters for $q(\pi)$ and $q(A)$. Then, using Jensen inequality, we derive the variational objective $\mathcal{L}$ called by evidence lower bound (ELBO) as

$$\log p(Y|X) \geq \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y, Z|X, A, \pi)\right] - KL(q(A, \pi)||p(A, \pi)) = \mathcal{L}, \tag{5}$$

where $\mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y, Z|X, A, \pi)\right]$ is expressed as

$$\mathbb{E}_{q(\pi)}\left[\log p(z_0|\pi)\right] + \mathbb{E}_{q(Z,A)}\left[\sum_{t=1}^{T} \log p(z_t|z_{t-1}, A)\right] + \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T} \log p(y_t|x_t, z_t)\right]. \tag{6}$$

For the training, $\mathcal{L}$ is maximized by alternatively updating the local hidden variables $q(Z)$, variational parameters w^π and $\{w_j^A\}_{j=1}^K$ for $q(\pi, A)$, and kernel hyperparameters $\{\theta_k\}_{k=1}^K$ based on variational EM algorithm. The detailed update procedures are described in the appendix.

2.3. Scalable Variational Inference

2.3.1. SVI APPROACH FOR LONG SEQUENCES (T IS LARGE)

Given the T sequences of time-series, we randomly sample L consecutive sequences of inputs and outputs $X_L^s = \{x_{k-1+l}\}_{l=1}^L$ and $Y_L^s = \{y_{k-1+l}\}_{l=1}^L$, where k is sampled uniformly from $k \in \{1, .., T-L+1\}$. We linearly approximate the ELBO of full sequences by considering the expected log joint likelihood of $\{Z_L^s, Y_L^s\}$ given X_L^s as

$$\begin{aligned}&\mathbb{E}_s\left[\mathbb{E}_q\left[\log p(Y_L^s, Z_L^s|X_L^s)\right]\right]\\ &\approx \frac{1}{T-L+1}\mathbb{E}_q\left[\sum_{t=1}^{T-L+1}\log p(z_{t-1}) + L\sum_{t=1}^{T}\log p(z_t|z_{t-1}, A) + L\sum_{t=1}^{T}\log p(y_t|z_t, x_t)\right]. \quad (7)\end{aligned}$$

Based on Eq. (7), the batch factors C_s^A and C_s^θ for calibrating the ELBO of $\{X_L^s, Y_L^s\}$ with full ELBO of $\{X, Y\}$ are obtained as

$$C_s^A = \frac{T-L+1}{L}\ , \qquad C_s^\theta = \frac{T-L+1}{L}. \quad (8)$$

The detailed derivation of batch factors are explained in appendix. With the batch factor C_s^A in Eq. (8) and the estimated $q^*(Z_L^s)$ for Y_L^s and X_L^s, the variational parameters w^π for $q(\pi)$ and w^A for $q(A)$ are updated by stochastic natural gradient descent as

$$w_j^\pi = (1-p_n)w_j^\pi + p_n\left(\alpha_j^\pi + q_s^*(z_k = j)\right) \quad (9)$$

$$w_{j,i}^A = (1-p_n)w_{j,i}^A + p_n\left(\alpha_{j,i}^A + C_s^A\sum_{l=1}^{L} q_s^*(z_{k-1+l} = j, z_{k+l} = i)\right), \quad (10)$$

where p_n is n-th iteration learning rate. SM kernel hyperparameters $\theta = \{\theta_k\}_{k=1}^K$ are updated by maximizing the following expected log marginal likelihood coordinated with the batch factor C_s^θ as

$$C_s^\theta\ \mathbb{E}_{q(Z_L^s)}\left[\sum_{l=0}^{L}\log p(y_{k-1+l}|z_{k-1+l}, x_{k-1+l})\right]. \quad (11)$$

2.3.2. APPROXIMATE GP EMISSION FOR A LARGE NUMBER OF TIME-SERIES (N_t IS LARGE)

For the SM kernel $k_{SM}(x-y) = \sum_{q=1}^Q w_q \exp\left(-2\pi^2((x-y)^T\sigma_q)^2\right)\cos\left(2\pi(x-y)^T\mu_q\right)$ with the parameters $\{w_q, \mu_q, \sigma_q^2\}_{q=1}^Q$, we approximate the SM kernel based on RFF.

Let $S = [S_{1,1}, .., S_{1,m}, .., S_{Q,1}, .., S_{Q,m}]$ be the random spectral points and $q(S)$ be the variational distribution of S defined as

$$q(S) = \prod_{q=1}^{Q}\prod_{i=1}^{m} N(S_{q,i}; \mu_q, \sigma_q^2). \quad (12)$$

Using the set of the spectral points $\boldsymbol{s} = \cup_{q=1}^Q\{s_{q,i}\}_{i=1}^m \sim q(S)$ sampled by reparametrization trick as $s_{q,i} = \mu_q + \sigma_q \circ \epsilon_i$ with $\epsilon_i \sim N(\epsilon; 0, I)$, we define the feature map $\phi^{SM}(x; \boldsymbol{s})$ as

$$\begin{aligned}\phi^{SM}(x; \boldsymbol{s}) &= \left[\sqrt{w_1}\phi_{\{s_{1,i}\}_{i=1}^m}(x), .., \sqrt{w_Q}\phi_{\{s_{Q,i}\}_{i=1}^m}(x)\right] \in R^{1\times 2Qm} \quad (13)\\ \phi_{\{s_{q,i}\}_{i=1}^m}(x) &= \frac{1}{m}\left[\cos\left(x_{\{s_{q,i}\}_{i=1}^m}\right), \sin\left(x_{\{s_{q,i}\}_{i=1}^m}\right)\right] \in R^{1\times 2m},\end{aligned}$$

to approximate the SM kernel $k_{SM}(x-y)$ by $\phi^{SM}(x;\boldsymbol{s})\phi^{SM}(y;\boldsymbol{s})^T$ and gram matrix $k_{SM}(x_t,x_t) \in R^{N_t\times N_t}$ by $\Phi^{SM}(x_t;\boldsymbol{s})\Phi^{SM}(x_t;\boldsymbol{s})^T$ with the feature matrix $\Phi^{SM}(x_t;\boldsymbol{s}) \in R^{N_t\times 2Qm}$.

Then, using Jensen's inequality, we derive the regularized lower bound of $\log p(y_t|z_t,x_t)$ as

$$\log p(y_t|x_t,z_t) \gtrapprox \frac{1}{K}\sum_{k=1}^{K} \log p(y_t|x_t,z_t,\boldsymbol{s}^{(k)}) - KL(q(S)||p(S)), \tag{14}$$

where $p(y_t|x_t,z_t,\boldsymbol{s}^{(k)})$ is computed as $N(y_t|m_{z_t}(x_t), \Phi^{SM}_{z_t}(x_t;\boldsymbol{s}^{(k)})\Phi^{SM}_{z_t}(x_t;\boldsymbol{s}^{(k)})^T + \sigma^2_{\epsilon_{z_t}} I)$ with the set of spectral points $\boldsymbol{s}^{(k)}$ sampled from $q(S)$ at k times. Eq (14) enables efficient computation because the inversion of the approximate gram matrix is efficiently computable (Lázaro-Gredilla et al., 2010). The prior distribution of $p(S)$ is assumed as $\prod_{q=1}^{Q}\prod_{i=1}^{m} N(S_{q,i};\widetilde{\mu}_{q,i},\widetilde{\sigma}^2_{q,i})$.

Applying the result of Eq. (14) to the conditional log likelihood $\log p(y_t|x_t,z_t)$ in Eq. (6), we derive the lower bound $\mathcal{L}_{asm}$, which can be efficiently computable, as

$$\begin{aligned}\mathcal{L}_{asm} = \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] - T\sum_{k=1}^{K}\sum_{q=1}^{Q} \mathrm{KL}(N(u_q,\sigma_q^2) \,||\, N(\widetilde{\mu}_{q,1},\widetilde{\sigma}^2_{q,1})) \\ - KL(q(A,\pi)||p(A,\pi)),\end{aligned} \tag{15}$$

under the condition that the parameters of prior $p(S)$ are assumed as $\{\widetilde{\mu}_{q.i},\widetilde{\sigma}^2_{q,i}\}_{i=2}^{M} = \{\mu_q,\sigma_q^2\}$. Here, $\mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right]$ is the expected joint likelihood over $q(Z,A,\pi)$ using the approximate GP emission $p(y_t|x_t,z_t)$ defined as

$$\prod_{k=1}^{K} N\left(y_t|m_{z_t}(x_t), \Phi^{SM}_{z_t}(x_t;\boldsymbol{s}^{(k)})\Phi^{SM}_{z_t}(x_t;\boldsymbol{s}^{(k)})^T + \sigma^2_{\epsilon_{z_t}} I\right)^{\frac{1}{K}}. \tag{16}$$

Optimizing the $\mathcal{L}_{asm}$ in Eq. (15) can be efficiently conducted in the same way as $\mathcal{L}$ in Eq. (5) is optimized, using the less number of operation for updating the parameters. It can be used together with section 2.3.1 SVI approach for long sequences dataset.

3. Experiments

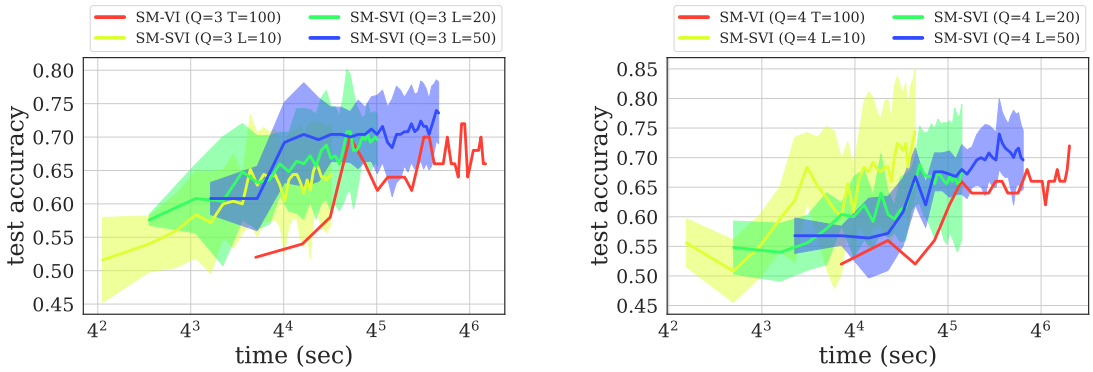
Evaluation Metric

• Accuracy : The ratio of correct hidden state estimation for sequences of time-series observations. Since the estimated hidden states $\{\hat{z}_t\}_{t=1}^{T}$ are possibly equivalent to the true hidden states up to the permutation of hidden states, we reorder them by Munkres algorithm (Munkres, 1957); For example, if two sequences of the estimated hidden states are given as $z^1_{1:3} = (1,2,3)$ and $z^2_{1:3} = (2,3,1)$, and there exists the correspondence π between hidden states $z^1_{1:3}$ and $z^2_{1:3}$ to satisfy $\pi(1) = 2, \pi(2) = 3$, and $\pi(3) = 1$, those estimated hidden states $z^1_{1:3}$ and $z^2_{1:3}$ are equivalent. Then, we calculate the ratio of how many reordered states are matched with the total T true hidden states.

• 1 iteration time : single iteration time (seconds) required to update local and global parameters.

Simulation data

We generate the sequences of time-series observations with eight different hidden states. We assume that state transition follows Markov process. Specifically, we construct two groups of hidden states whose dynamics are different: $(1 \to 2 \to 3 \to 4)$ and $(5 \to 6 \to 7 \to 8)$. In the first group, each



(a) Test set accuracy (Q=3)

(b) Test set accuracy (Q=4)

Figure 1: Comparison of training HMM-GPSM ($Q \in \{3, 4\}$) by VI ($T = 100$) and SVI ($L \in \{10, 20, 50\}$) for $W_{Hz} = 100$

state except state 4 follows the staying probability 0.7 and the moving probability 0.3. Similarly, in the second group, each state except state 8 has staying probability 0.3 and the moving probability 0.7. The special states $\{4, 8\}$ connect the two group with probability 1, i.e., $(4 \to 5)$ and $(8 \to 1)$. Given the hidden state $s \in \{1, .., 8\}$, the time-series observation $f^s(t)$ is generated by

$$f^s(t) = \sum_{q=1}^{6} \alpha_q^s \sin\left(2\pi\, \omega_q^s\, t\right) + \epsilon \tag{17}$$

where the weights $\alpha^s = [\alpha_1^s, .., \alpha_6^s]$ are sampled from $\mathrm{Dir}(\alpha^s|\alpha_0)$, and the frequency components ω_q^s are sampled from the interval $[0, W_{Hz}]$.

3.1. Long Sequences (T is large)

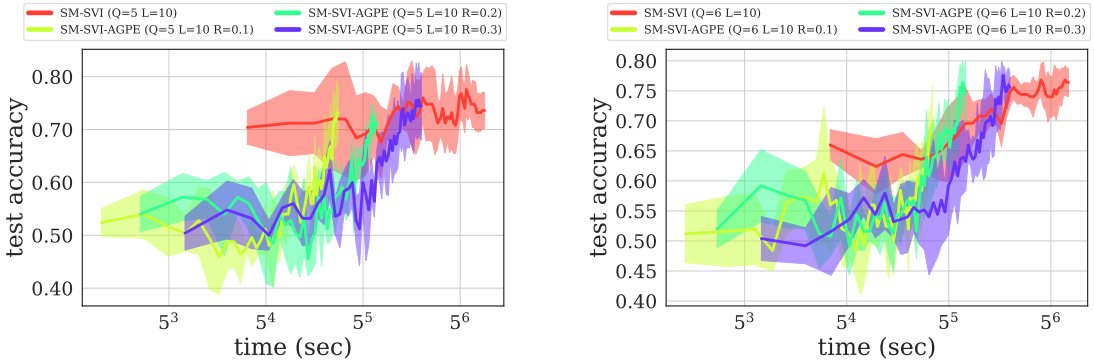
We investigate how the SVI approach reduces the training time required to train HMM-GPSM with long sequences of time-series dataset (T is large). We compare the following inference methods:

- VI (T) : Variational inference for full T sequences in appendix
- SVI (L) : Stochastic Variational inference with L sampled sub-sequences in section 2.3.1

We generate the sequences of time-series observations using Eq. (17) having 150 sequences of time-series. We fix the number of data points in each time-series observation as $N_t = 200$. We limit spectral range as $W_{Hz} = 100$ to make sure that time-series with $N_t = 200$ contains enough information to infer the spectral characteristics of the time-series according to the Nyquist–Shannon sampling theorem (Oppenheim, 1999).

We train the first 150 sequences of times-series ($T = 100, N_t = 200$) and test the left sequences ($T = 50, N_t = 200$); VI approach uses the whole sequences ($T = 100$) of the training data, while the SVI approach uses the sub-sequences with the length $L \in \{10, 20, 50\}$. In addition, to see how the model complexity affects the training time, we consider the number of mixture for the SM kernel as $Q \in \{3, 4\}$.

Figure 1(a) compares the training performance of the SVI training approach for varying batch length $L \in \{10, 20, 50\}$ with the fixed $Q = 3$. Figure 1(b) shows the same results for $Q = 4$.



(a) Test set accuracy (Q=5)

(b) Test set accuracy (Q=6)

Figure 2: Comparison of training HMM-GPSM ($Q \in \{5, 6\}$) by SVI and SVI-AGPE ($L = 10$ and $R \in \{.10, .20, .30\}$) for $W_{Hz} = 500$

In this plot, each test accuracy for VI and SVI is evaluated every iteration during total 30 training iterations. Figures show that SVI approach with the smaller batch length $L \in \{10, 20, 50\}$ takes less training time to achieve a similar level of test accuracy compared to the VI approach.

3.2. Large number of observations (N_t is large)

In this experiment, we investigate how the derived ELBO $\mathcal{L}_{asm}$ in Eq. (15) reduces the training time of the HMM-GPSM and affects the accuracy of hidden state estimation when N_t is large. To this end, we compare the following two inference methods:

- SVI (L) : Stochastic Variational inference with L sampled sub-sequences in section 2.3.1.
- SVI-AGPE (L, R) : Stochastic Variational inference with L sampled sub-sequences in section 2.3.1. and approximate GP emission with the sampling rate R in section 2.3.2.

During the experiment, we change the number of mixtures Q for SM kernel and the ratio of the sampled spectral points R used for the SM kernel approximation to investigate how these parameters affect the accuracy and training time; the number of sampled spectral points M is set as $M \propto N_t R$.

We generate 150 sequences of time-series observation by Eq. (17) with $N_t = 1000$ and $W_{Hz} = 500$. The first 100 sequences ($T = 100, N_t = 1000$) and the left 50 sequences ($T = 50, N_t = 1000$) are used for training and test, respectively. For the number of repetition K for SM kernel approximation in Eq. (16), we use $K = 1$ for training and $K = 3$ for test.

Figure 2(a) compares the training performance of the SVI-AGPE approach for varying sampling rates $R \in \{.10, .20, .30\}$ with the fixed $L = 10$ and $Q = 5$. Figure 2(b) shows the same results for $Q = 6$. Each test accuracy for SVI and SVI-AGPE approaches is evaluated every iteration during 50 training iterations. Figures show that SVI-AGPE with the rate $R \in \{.10, .20, .30\}$ takes less training time to achieve a similar level of test accuracy for SVI.

4. Conclusion

In this work, we propose the scalable learning method for HMM-GPSM and validate that the proposed learning algorithm scalably trains HMM-GPSM with a large-scale dataset while maintaining the clustering performance.

References

Matthew J Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, UCL (University College London), 2003.

Yoshua Bengio and Paolo Frasconi. An input output hmm architecture. In *Advances in neural information processing systems*, pages 427–434, 1995.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Salomon Bochner. *Lectures on Fourier integrals*. Princeton University Press, 1959.

Nick Foti, Jason Xu, Dillon Laird, and Emily Fox. Stochastic variational inference for hidden markov models. In *Advances in neural information processing systems*, pages 3599–3607, 2014.

Roger Frigola. *Bayesian time series learning with Gaussian processes*. PhD thesis, University of Cambridge, 2015.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl Edward Rasmussen, and Aníbal R Figueiras-Vidal. Sparse spectrum gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881, 2010.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

Masatoshi Nagano, Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, and Masahide Kaneko. Sequence pattern extraction by segmenting time series data using gp-hsmm with hierarchical dirichlet process. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4067–4074. IEEE, 2018.

Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, Hideki Asoh, and Masahide Kaneko. Segmenting continuous motions with hidden semi-markov models and gaussian processes. *Frontiers in neurorobotics*, 11:67, 2017.

Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

Kostantinos N Plataniotis and Dimitris Hatzinakos. Gaussian mixtures and their applications to signal processing. In *Advanced signal processing handbook*, pages 89–124. CRC Press, 2017.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

Kyle R Ulrich, David E Carlson, Wenzhao Lian, Jana S Borg, Kafui Dzirasa, and Lawrence Carin. Analysis of brain states from multi-region lfp time-series. In *Advances in Neural Information Processing Systems*, pages 2483–2491, 2014.

Kyle R Ulrich, David E Carlson, Kafui Dzirasa, and Lawrence Carin. Gp kernels for cross-spectrum analysis. In *Advances in neural information processing systems*, pages 1999–2007, 2015.

Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

SUPPLEMENTARY MATERIAL

Appendix A. Derivation

A.1. ELBO Derivation for Eq. (5)

$$
\begin{aligned}
&\log p(Y|X) \\
&= \iiint p(Y|X,Z,A,\pi)p(Z,A,\pi|X)\; dZdAd\pi \\
&\geq \iiint \log\Big(p(Y|X,Z,A,\pi)\frac{p(Z,A,\pi)}{q(Z,A,\pi)}\Big)q(Z,A,\pi)\; dZdAd\pi \\
&= \iiint \log p(Y|X,Z,A,\pi)q(Z,A,\pi)\; dZdAd\pi - KL(q(Z,A,\pi)||p(Z,A,\pi)) \\
&= \iiint \log p(Y,Z|X,A,\pi)q(Z,A,\pi)\; dZdAd\pi + H(q(Z),p(Z)) - KL(q(Z,A,\pi)||p(Z,A,\pi)) \\
&= \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] + H(q(Z)) - KL(q(A,\pi)||p(A,\pi)) \\
&\geq \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] - KL(q(A,\pi)||p(A,\pi)) \coloneqq \mathcal{L}
\end{aligned}
$$

where cross entropy $H(q(Z),p(Z)) = H(q(Z)) + KL(q(Z)||p(Z))$.

A.2. Batch factor Derivation for SVI Approach for Eq. (8)

Given the sampled $i \in \{1,..,T-L+1\}$ uniformly, let $Y_L^s = \{y_i,..,y_{i+L-1}\}$ be sampled observation with the length L and X_L^s be corresponding inputs and Z_L^s corresponding hidden states. The expected log joint likelihood of $\{Z_L^s, Y_L^s\}$ given X_L^s is approximated as

$$
\begin{aligned}
&\mathbb{E}_s\Big[\mathbb{E}_q[\log p(Y_L^s, Z_L^s|X_L^s)]\Big] \\
&= \sum_{i=0}^{T-L}\frac{1}{T-L+1}\mathbb{E}_q[\log p(Y_L^{s_i}, Z_L^{s_i}|X_L^{s_i})] \\
&= \frac{1}{T-L+1}\sum_{i=0}^{T-L}\mathbb{E}_q\left[\log p(z_i) + \underbrace{\sum_{t=1}^{L}\log p(z_{i+t}|z_{i+t-1},A)}_{\text{transition term}} + \underbrace{\sum_{t=1}^{L}\log p(y_{i+t}|z_{i+t},x_{i+t})}_{\text{observation term}}\right] \cdots (*) \\
&\approx \frac{1}{T-L+1}\mathbb{E}_q\left[\sum_{t=1}^{T-L+1}\log p(z_{t-1}) + L\sum_{t=1}^{T}\log p(z_t|z_{t-1},A) + L\sum_{t=1}^{T}\log p(y_t|z_t,x_t)\right]
\end{aligned}
$$

This implies that transition and observation term of $\mathbb{E}_q[\log p(Y_L^s, Z_L^s|X_L^s)]$ for the sampled $i \in \{1,..,T-L+1\}$ can be approximated as

$$
\begin{aligned}
\mathbb{E}_q\left[\sum_{t=1}^{L}\log p(z_{i+t}|z_{i+t-1},A)\right] &\approx \frac{L}{T-L+1}\mathbb{E}_q\left[\sum_{t=1}^{T}\log p(z_t|z_{t-1},A)\right] \\
\mathbb{E}_q\left[\sum_{t=1}^{L}\log p(y_{i+t}|z_{i+t},x_{i+t})\right] &\approx \frac{L}{T-L+1}\mathbb{E}_q\left[\sum_{t=1}^{T}\log p(y_t|z_t,x_t)\right]
\end{aligned}
$$

Thus, the batch factors, C_s^A and C_s^θ, to calibrate the approximated ELBO are obtained as

$$C_s^A = \frac{T-L+1}{L} \ , \ C_s^\theta = \frac{T-L+1}{L}$$

The transition term in expectation in $(*)$ can be approximated as

$$\begin{aligned}
&\sum_{i=0}^{T-L} \mathbb{E}_q \left[\sum_{t=1}^{L} \log p(z_{i+t}|z_{i+t-1}, A) \right] \\
&= \mathbb{E}_q \left[\sum_{j=1}^{L} \log p(z_j|z_{j-1}, A) + \sum_{j=2}^{L+1} \log p(z_j|z_{j-1}, A) + \cdots + \sum_{j=T-L+1}^{T} \log p(z_j|z_{j-1}, A) \right] \\
&= \mathbb{E}_q \left[L \sum_{t=L}^{T-L+1} \log p(z_t|z_{t-1}, A) + \underbrace{\sum_{t=1}^{L-1} t\Big(\log p(z_t|z_{t-1}, A) + \log p(z_{T-t+1}|z_{T-t}, A) \Big)}_{\text{approximate term}} \right] \\
&\approx \mathbb{E}_q \left[L \sum_{t=L}^{T-L+1} \log p(z_t|z_{t-1}, A) + L \sum_{t=1}^{L-1} \Big(\log p(z_t|z_{t-1}, A) + \log p(z_{T-t+1}|z_{T-t}, A) \Big) \right] \\
&= \mathbb{E}_q \left[L \sum_{t=1}^{T} \log p(z_t|z_{t-1}, A) \right]
\end{aligned}$$

Here, the observation term in expectation in $(*)$ can be approximated as

$$\begin{aligned}
&\sum_{i=0}^{T-L} \mathbb{E}_q \left[\sum_{t=1}^{L} \log p(y_{i+t}|z_{i+t}, x_{i+t}) \right] \\
&= \mathbb{E}_q \left[\sum_{j=1}^{L} \log p(y_j|z_j, x_j) + \cdots + \sum_{j=T-L+1}^{T} \log p(y_j|z_j, x_j) \right] \\
&= \mathbb{E}_q \left[L \sum_{t=L}^{T-L+1} \log p(y_t|z_t, x_t) + \underbrace{\sum_{t=1}^{L-1} t\Big(\log p(y_t|z_t, x_t) + \log p(y_{T-t+1}|z_{T-t+1}, x_{T-t+1}) \Big)}_{\text{approximate term}} \right] \\
&\approx \mathbb{E}_q \left[L \sum_{t=L}^{T-L+1} \log p(y_t|z_t, x_t) + L \sum_{t=1}^{L-1} \Big(\log p(y_t|z_t, x_t) + \log p(y_{T-t+1}|z_{T-t+1}, x_{T-t+1}) \Big) \right] \\
&= \mathbb{E}_q \left[L \sum_{t=1}^{T} \log p(y_t|z_t, x_t) \right]
\end{aligned}$$

A.3. SM kernel Approximation for Eq. (13)

Given the parameters of SM kernel $\{w_q, \mu_q, \sigma_q\}_{q=1}^{Q}$, we sample spectral points $\boldsymbol{s}_q = \{s_{q,i}\}_{i=1}^{m}$ from Gaussian distribution $N(S; \mu_q, \sigma_q)$ by reparametrization trick as

$$s_{q,i} = \mu_q + \sigma_q \circ \epsilon_i$$

where $\epsilon_i \sim N(\epsilon; 0, I)$ for $i = 1, .., m$. If we define the feature map $\phi_{\boldsymbol{s}_q}(x)$ as

$$\phi_{\boldsymbol{s}_q}(x) = \frac{1}{\sqrt{m}} [\cos 2\pi s_{q,1}, \sin 2\pi s_{q,1}, .., \cos 2\pi s_{q,m}, \sin 2\pi s_{q,m}] \in R^{1\times 2m}$$

, then $\phi_{\boldsymbol{s}_q}(x)\phi_{\boldsymbol{s}_q}(y)^T$ can approximate $k_q(x-y)$ which is the inducted kernel from Gaussian Spectral density $N(S; \mu_q, \sigma_q)$ by Bochner's theorem.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)} \left[\phi_{\boldsymbol{s}_q}(x)\phi_{\boldsymbol{s}_q}(y)^T\right] \\
&= \mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)} \left[\frac{1}{m_q}\sum_{i=1}^{m_q} \left(\cos 2\pi {s_{q,i}}^T x\right)\left(\cos 2\pi {s_{q,i}}^T y\right) + \left(\sin 2\pi {s_{q,i}}^T x\right)\left(\sin 2\pi {s_{q,i}}^T y\right)\right] \\
&= \mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)} \left[\frac{1}{m_q}\sum_{i=1}^{m_q} \cos 2\pi {s_{q,i}}^T (x-y)\right] \\
&= \mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)} \left[\frac{1}{m_q}\sum_{i=1}^{m_q} \frac{e^{i2\pi {s_{q,i}}^T (x-y)} + e^{-i2\pi {s_{q,i}}^T (x-y)}}{2}\right] \\
&= \frac{1}{2}\left(k_q(x-y) + k_q(y-x)\right) = k_q(x-y)
\end{aligned}$$

Using the above derivation, if we define sampled spectral points $\boldsymbol{s} = \cup_{q=1}^{Q}\{s_{q,i}\}_{i=1}^{m}$ with $s_{q,i} \sim N(\mu_q, \sigma_q^2)$ and the feature map $\phi^{SM}(x) = \left[\sqrt{w_1}\phi_{\{s_{1,i}\}_{i=1}^{m}}(x), .., \sqrt{w_Q}\phi_{\{s_{Q,i}\}_{i=1}^{m}}(x)\right]$, then $\phi^{SM}(x)\phi^{SM}(y)^T$ is an unbiased estimator of $k_{SM}(x, y)$ as

$$\mathbb{E}_{\boldsymbol{s}}\left[\phi^{SM}(x)\phi^{SM}(y)^T\right] = \sum_{q=1}^{Q} w_q \mathbb{E}_{\boldsymbol{s}_q \sim N(S;\mu_q,\sigma_q)}\left[\phi_{\boldsymbol{s}_q}(x)\phi_{\boldsymbol{s}_q}(y)^T\right] = \sum_{q=1}^{Q} w_q k_q(x-y) = k_{SM}(x-y)$$

A.4. Regularized Lower bound for Eq. (14)

Let $q(S)$ be variational distribution defined in Eq. (12). We can derive the lower bound $\mathcal{L}$ as follows:

$$\begin{aligned}
\log p(Y|X) &= \log \int p(Y, S|X) dS = \log \int p(Y|X, S)\frac{p(S)}{q(S)} q(S) dS \\
&\geq \int \log\left(p(Y|X, S)\frac{p(S)}{q(S)}\right) q(S) dS \\
&= \int \log p(Y|X, S) q(S) + \log \frac{p(S)}{q(S)} q(S) dS \\
&= \int \log p(Y|X, S) q(S) dS - KL(q(S)||P(S)) \\
&\approx \frac{1}{K}\sum_{i=1}^{K} \log p(Y|X, s^{(i)}) - KL(q(S)||P(S))
\end{aligned}$$

where $s^{(i)}$ is i-th sampled spectral points from $q(S)$.

A.5. ELBO Derivation for Eq. (15)

$$
\begin{aligned}
\mathcal{L} &= \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] - KL(q(A,\pi)||p(A,\pi)) \\
&= \mathbb{E}_{q(\pi)}\left[\log p(z_0|\pi)\right] + \mathbb{E}_{q(Z,A)}\left[\sum_{t=1}^{T}\log p(z_t|z_{t-1},A)\right] + \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T}\log p(y_t|x_t,z_t)\right] \\
&\qquad - KL(q(A,\pi)||p(A,\pi)) \\
&\geq \mathbb{E}_{q(\pi)}\left[\log p(z_0|\pi)\right] + \mathbb{E}_{q(Z,A)}\left[\sum_{t=1}^{T}\log p(z_t|z_{t-1},A)\right] + \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T}\frac{1}{K}\sum_{k=1}^{K}\log p(y_t|z_t,x_t,\boldsymbol{s}^{(k)})\right] \\
&\qquad - \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T}KL(q(S|z_t)||p(S|z_t))\right] - KL(q(A,\pi)||p(A,\pi)) \\
&\geq \mathbb{E}_{q(\pi)}\left[\log p(z_0|\pi)\right] + \mathbb{E}_{q(Z,A)}\left[\sum_{t=1}^{T}\log p(z_t|z_{t-1},A)\right] + \mathbb{E}_{q(Z)}\left[\sum_{t=1}^{T}\frac{1}{K}\sum_{k=1}^{K}\log p(y_t|z_t,x_t,\boldsymbol{s}^{(k)})\right] \\
&\qquad - T\sum_{k=1}^{K}KL(q(S|z_t=k)||p(S|z_t=k)) - KL(q(A,\pi)||p(A,\pi))
\end{aligned}
$$

The first inequality holds by applying the lower bound bound in Eq. (14) to $\log p(y_t|x_t,z_t)$ in the derived ELBO in Eq. (5). The second inequality is derived using $q(z_t=k)\leq 1$ for all t,k.

$$
= \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] - T\sum_{k=1}^{K}KL(q(S|z_t=k)||p(S|z_t=k)) - KL(q(A,\pi)||p(A,\pi))
$$

where $\mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right]$ is the expected joint likelihood with respect to $q(Z,A,\pi)$ using $\frac{1}{K}\sum_{k=1}^{K}\log p(y_t|z_t,x_t,\boldsymbol{s}^{(k)})$ as log likelihood of GP emission instead of $\log p(y_t|z_t,x_t)$.

The additional KL divergence term related to $KL(q(S|z_t=k)||p(S|z_t=k))$ is computed as $\sum_{q=1}^{Q}\sum_{i=1}^{m}\mathrm{KL}(N(u_q,\sigma_q^2)||N(\widetilde{\mu}_{q,i},\widetilde{\sigma}_{q,i}^2))$ with the parameters $\{\mu_q,\sigma_q\}_{q=1}^{Q}$ corresponding to hidden state $z_t=k$. If we consider $\{\widetilde{\mu}_{q.i},\widetilde{\sigma}_{q,i}^2\}_{i=2}^{M}=\{\mu_q,\sigma_q\}$, we reduce the KL regulaizer as $\sum_{q=1}^{Q}\mathrm{KL}(N(u_q,\sigma_q^2)||N(\widetilde{\mu}_{q,1},\widetilde{\sigma}_{q,1}^2))$.

$$
= \mathbb{E}_{q(Z,A,\pi)}\left[\log p(Y,Z|X,A,\pi)\right] - T\sum_{k=1}^{K}\sum_{q=1}^{Q}\mathrm{KL}(N(u_q,\sigma_q^2)\;||\;N(\widetilde{\mu}_{q,1},\widetilde{\sigma}_{q,1}^2)) - KL(q(A,\pi)||p(A,\pi))
$$

For the corresponding approximate GP emission, we can obtain $\prod_{k=1}^{K}p(y_t|z_t,x_t,\boldsymbol{s}^{(k)})^{\frac{1}{k}}$ because of

$$
\log p(y_t|z_t,x_t) \approx \frac{1}{K}\sum_{k=1}^{K}\log p(y_t|z_t,x_t,\boldsymbol{s}^{(k)}) = \log\prod_{k=1}^{K}p(y_t|z_t,x_t,\boldsymbol{s}^{(k)})^{\frac{1}{k}}.
$$

If $\{\boldsymbol{s}^{(k)}\}_{k=1}^{K}$ are sampled such that $p(y_t|z_t, x_t, \boldsymbol{s}^{(k)}) \approx p(y_t|z_t, x_t)$, the approximate GP emission $\prod_{k=1}^{K} p(y_t|z_t, x_t, \boldsymbol{s}^{(k)})^{\frac{1}{k}}$ becomes the original GP emission $p(y_t|z_t, x_t)$. This can be feasible when the number of sampled spectral points Qm for $\boldsymbol{s}^{(k)}$ is large enough for $\Phi^{SM}(x_t; \boldsymbol{s}^{(k)})\Phi^{SM}(x_t; \boldsymbol{s}^{(k)})^T$ to be equal to the true SM kernel gram matrix $K_{SM}(x_t, x_t)$.

Appendix B. Parameters update procedure for section 2.2 Variational Inference

B.1. Local Variables

The optimal local variable $q^*(Z)$ is proportional as

$$\begin{aligned} q^*(Z) &\propto \exp\left(\mathbb{E}_{q(A,\pi)}\left[\log p(Y, Z|X, A, \pi)\right]\right) \\ &= \exp\left(\mathbb{E}_{q(\pi)}[\log p(z_0|\pi)] + \sum_{t=1}^{T} \mathbb{E}_{q(A)}[\log p(z_t|z_{t-1}, A)] + \sum_{t=1}^{T} \log p(y_t|z_t, x_t)\right), \end{aligned} \tag{18}$$

by Mean field approximation (Bishop, 2006; Beal, 2003). We define the following auxiliary variables $\widetilde{\pi}$ as $\exp\left(\mathrm{E}_{q(\pi)}[\log \pi_j]\right)$ and $\widetilde{A}_{j,i}$ as $\exp\left(\mathrm{E}_{q(A)}[\log A_{j,i}]\right)$ required to evaluate Eq. (18), which are computed as

$$\begin{aligned} \widetilde{\pi} &:= \exp\left(\mathrm{E}_{q(\pi)}[\log \pi_j]\right) = \exp\left[\psi(w_j^{\pi}) - \psi\left(\sum_{j=1}^{K} w_j^{\pi}\right)\right] \\ \widetilde{A}_{j,i} &:= \exp\left(\mathrm{E}_{q(A)}[\log A_{j,i}]\right) = \exp\left[\psi(w_{j,i}^{A}) - \psi\left(\sum_{i=1}^{K} w_{j,i}^{A}\right)\right] \end{aligned}$$

where $\psi(\cdot)$ is the digamma function.

To compute the marginal distribution $q^*(z_t = k)$ and $q^*(z_t = j, z_t = i)$ necessary for updating variational parameters for global variables, forward-backward algorithm known as Baum–Welch algorithm (Beal, 2003) is used. Defining $\alpha_{t,i} = p(z_t = i|y_{1:t}, x_{1:t})$ and $\beta_{t,i} = p(y_{t+1:T}|z_t = i, x_{t+1:T})$ with $\alpha_0 = \pi$ and $\beta_T = [1, .., 1]^T \in R^K$, we compute $\{\alpha_t, \beta_t\}_{t=1}^{T}$ as

$$\begin{aligned} \alpha_{t,i} &= \sum_{j=1}^{K} \alpha_{t-1,j} \widetilde{A}_{j,i} p\left(y_t|z_t = i, x_t\right) \\ \beta_{t,j} &= \sum_{i=1}^{K} \widetilde{A}_{j,i} p\left(y_{t+1}|z_{t+1} = i, x_{t+1}\right) \beta_{t+1,i}. \end{aligned}$$

These computed likelihoods $\{\alpha_t, \beta_t\}_{t=1}^{T}$ are used to compute $q^*(z_t = i)$ and $q^*(z_t = j, z_t = i)$ for $t = 1, .., T$ and $j, i = 1, .., K$ as

$$\begin{aligned} q^*(z_t = i) &\propto \alpha_{t,i}\beta_{t,i} \\ q^*(z_{t-1} = j, z_t = i) &\propto \alpha_{t-1,j} A_{j,i}\, p(y_t|z_t = i, x_t)\beta_{t,i}. \end{aligned}$$

A detail derivation for Forward-backward algorithm can be found in (Beal, 2003).

Appendix C. Computational Complexity

To analyze the computation complexity for the proposed learning algorithm, we split the algorithm mainly into three parts; computation of log marginal likelihood for observations, local update, and global update. We proceed with the analysis of our computation under a single batch assumption because the repetitive batch sampling for SVI and the spectral points sampling of Eq. (14) increases the total computation linearly.

For the brevity, we assume $N_t = N$ for all t. In original VI approach, computing the log marginal likelihood of $T \times N$ observations with K hidden state costs $\mathcal{O}(KTN^3)$. However, our approximation approach costs $\mathcal{O}(KLNM^2)$ where the length of the sampled sequence is L, and the total M sampled spectral points are used for SM kernel approximation. This is because SVI approach reduces to $\mathcal{O}(L)$ from $\mathcal{O}(T)$ and SM kernel approximation reduces to $\mathcal{O}(NM^2)$ from $\mathcal{O}(N^3)$ under $M \ll N$.

For the update of local variables, VI and SVI take $\mathcal{O}(K^2T)$ and $\mathcal{O}(K^2L)$ for the Forward-Backward algorithm, respectively. Updating the global variable is dominated by updating kernel hyperparameters. Computing the derivative of log marginal likelihood for each parameter costs $\mathcal{O}(N^3)$ (Rasmussen, 2004). Thus, in the case of SM kernel, VI approach costs $\mathcal{O}((3Q+1)N^3KT)$ where all SM kernels take Q Gaussian mixture components. However, our scalable approach takes $\mathcal{O}((3Q+1)NM^2KL)$.

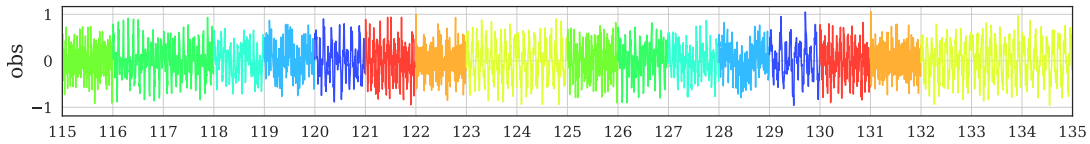
In summary, our learning method scalably trains the large dataset when we control M and L such that $NM^2 \ll N^3$ and $L \ll T$.

Appendix D. Additional results of Experiment

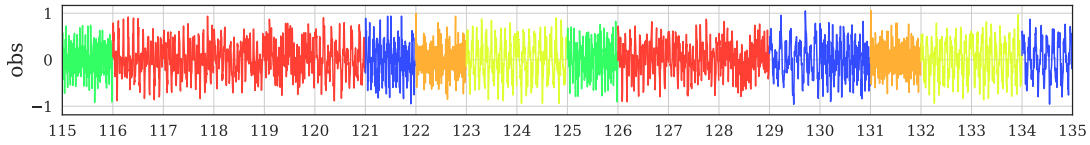
D.1. Long Sequence (T is large)

		$W_{Hz} = 100$	
Model	**Inference**	**Accuracy**	**1 Iteration**
HMM-GPSM (Q=3)	VI (T=100)	.66	173.18
HMM-GPSM (Q=3)	SVI (L=10)	.64 $\pm$.08	17.29 $\pm$.02
HMM-GPSM (Q=3)	SVI (L=20)	.70 $\pm$.04	34.46 $\pm$.13
HMM-GPSM (Q=3)	SVI (L=50)	.74 $\pm$.05	85.96 $\pm$.11
HMM-GPSM (Q=4)	VI (T=100)	.72	208.41
HMM-GPSM (Q=4)	SVI (L=10)	.74 $\pm$.11	20.94 $\pm$.02
HMM-GPSM (Q=4)	SVI (L=20)	.68 $\pm$.11	41.97 $\pm$.05
HMM-GPSM (Q=4)	SVI (L=50)	.70 $\pm$.05	104.57 $\pm$.19

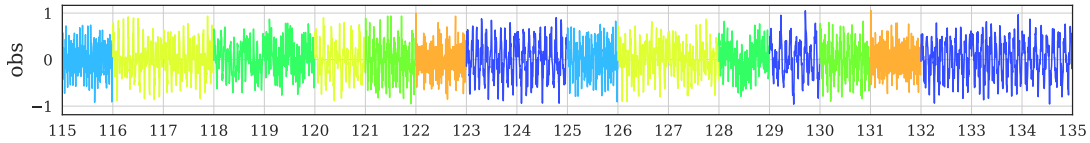
Table 1: Accuracy and single iteration time (seconds) on synthetic dataset with $W_{Hz} = 100$; This table describes the statistic of accuracy and single iteration time for training approach of VI and SVI with $Q \in \{3, 4\}$ and $L \in \{10, 20, 50\}$.



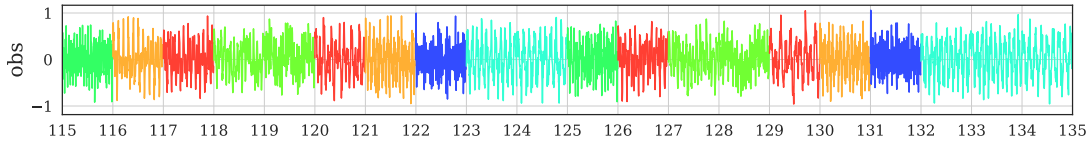
(a) Test data set



(b) HMM-GPSM (Q=4) trained by VI (T=100) with accuracy 0.72



(c) HMM-GPSM (Q=4) trained by SVI (L=10) with accuracy 0.84



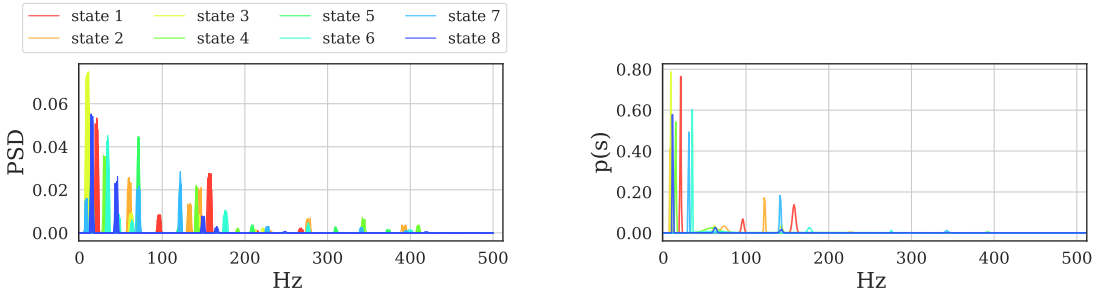
(d) HMM-GPSM (Q=4) trained by SVI (L=20) with accuracy 0.78

Figure 3: Test set clustering results of HMM-GPSM (Q=4) trained by VI (T=100) and SVI (L$\in \{10, 20\}$) for $W_{Hz} = 100$. We display the estimated result $\{\hat{z}_t\}_{t=115}^{134}$ on $[115, 135]$ out of $[100, 150]$ because of limited space. Each hidden state is characterized by its own color; if two time-series observations are colored in the same color, they are estimated to have same hidden state.

D.2. Large number of observations (N_t is large)

		$W_{Hz} = 500$	
Model	**Inference**	**Accuracy**	**1 Iteration**
HMM-GPSM (Q=3)	SVI (L=10)	.66 ± .02	325.29 ± 1.79
HMM-GPSM (Q=3)	SVI-AGPE (L=10 R=.05)	.64 ± .09	23.89 ± .19
HMM-GPSM (Q=3)	SVI-AGPE (L=10 R=.10)	.60 ± .03	32.43 ± .03
HMM-GPSM (Q=3)	SVI-AGPE (L=10 R=.20)	.64 ± .06	75.05 ± .38
HMM-GPSM (Q=5)	SVI (L=10)	.74 ± .03	468.81 ± 9.69
HMM-GPSM (Q=5)	SVI-AGPE (L=10 R=.10)	.73 ± .05	41.42 ± .10
HMM-GPSM (Q=5)	SVI-AGPE (L=10 R=.20)	.71 ± .02	78.38 ± .07
HMM-GPSM (Q=5)	SVI-AGPE (L=10 R=.30)	.74 ± .06	163.94 ± 2.02
HMM-GPSM (Q=6)	SVI (L=10)	.76 ± .02	515.54 ± 12.80
HMM-GPSM (Q=6)	SVI-AGPE (L=10 R=.10)	.70 ± .06	48.59 ± .28
HMM-GPSM (Q=6)	SVI-AGPE (L=10 R=.20)	.74 ± .06	81.38 ± .59
HMM-GPSM (Q=6)	SVI-AGPE (L=10 R=.30)	.76 ± .02	163.60 ± 1.63

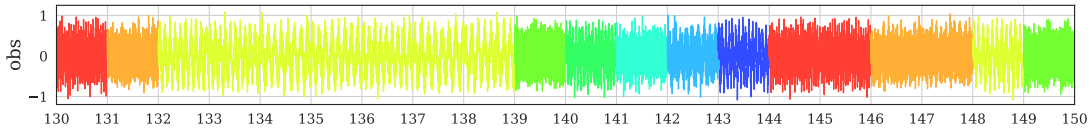
Table 2: Accuracy and single iteration time (seconds) on sinusoidal dataset with $W_{Hz} = 500$; This table summarizes how the proposed kernel approximation approach (SVI-AGPE) affects the training time and the accuracy for HMM-GPSM with $Q \in \{3, 5, 6\}$, comparing to the SVI approach without kernel approximation (SVI). For this comparison, we set $L = 10$ for SVI approach and vary the ratio of spectral points $R \in \{.05, .10, .20, .30\}$. The statistical results are obtained from 5 repetitive experiments.



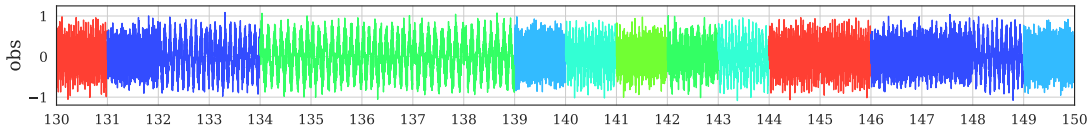
(a) Empirical spectral density

(b) Estimated spectral density

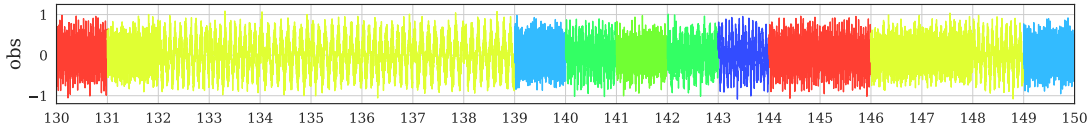
Figure 4: Comparison of empirical spectral density of time-series observation and estimated spectral density of HMM-GPSM (Q=5) trained by SVI-AGPE ($L = 10$ and $R = .20$) for $W_{Hz} = 500$; empirical spectral density is obtained by applying welch method (Welch, 1967) to 20 time-series of training set per each hidden state. Estimated spectral density is obtained by evaluating $p(s)$ in Eq. (5) with the parameters $\{w_q, \mu_q, \sigma_q^2\}_{q=1}^5$ in Eq. (30) inferred by SVI-AGPE.



(a) Test data set



(b) HMM-GPSM (Q=5) trained by SVI (L=10) with accuracy 0.74



(c) HMM-GPSM (Q=5) trained by SVI-AGPE (L=10, R=.20) with accuracy 0.78

Figure 5: Test set clustering results of HMM-GPSM (Q=5) trained by SVI (L=10) and SVI-AGPE (L=10, R $\in \{.10, .20\}$) for $W_{Hz} = 500$. We display the estimated result $\{\hat{z}_t\}_{t=130}^{149}$ on $[130, 150]$ out of $[100, 150]$ because of limited space.

D.3. Why is SM kernel considered

SM kernel is a flexible kernel to model any stationary covariance function. Bochner's theorem (Bochner, 1959) states that for the inputs x_1 and $x_2 \in R^P$ and its distance $\tau = x_1 - x_2$, the stationary kernel $k(\tau)$, which is invariant to translation of the inputs, can be obtained by taking a inverse Fourier transform to the corresponding spectral density of $p(S)$ as

$$k(\tau) = \int e^{i2\pi s^{\mathrm{T}}\tau} p(S) ds. \tag{19}$$

This theorem implies that if $p(S)$ approximates well the true spectral density for dataset, the target function f using GP prior with the corresponding kernel $k(\tau)$ models well the true signal.

Wilson et al. (Wilson and Adams, 2013) implement this idea to devise a new kernel known as the spectral mixture (SM) kernel. They represent the spectral density $p(S)$ using a mixture of symmetric Gaussian distribution with parameters $\{w_q, \mu_q, \Sigma_q\}_{q=1}^{Q}$ as

$$p(S) = \sum_{q=1}^{Q} w_q \left(\frac{N(s|\mu_q, \Sigma_q) + N(-s|\mu_q, \Sigma_q)}{2} \right),$$

where $\mu_q = [\mu_{q,1}, .., \mu_{q,P}], \sigma_q = [\sigma_{q,1}, .., \sigma_{q,P}] \in R^P$ and $\Sigma_q = \mathrm{diag}(\sigma_q^2) \in R^{P \times P}$ because a mixture of Gaussian distribution can approximate any continuous function by universal approximate theorem (Plataniotis and Hatzinakos, 2017). Then, they obtain the SM kernel by taking an inverse Fourier transform to $p(S)$ by Eq. (20) as

$$k_{SM}(\tau) = \sum_{q=1}^{Q} w_q \exp\left(-2\pi^2(\tau^T \sigma_q)^2\right) \cos\left(2\pi\tau^T \mu_q\right). \tag{20}$$

Flexibility of GP emission using SM kernel

We have investigated that GP emission using SM kernel characterizes the hidden state of time-series observation more flexibly over GP emission using the conventional kernel including RBF, Periodic, and the combination kernel. The considered baseline models are as follows:

- HMM-GPRBF : HMM-GP using RBF kernel
- HMM-GPPER : HMM-GP using Periodic kernel
- HMM-GPCombKernel : HMM-GP using combination kernel (RBF + PER + PER + PER)
- HMM-GPSM (Q) : HMM-GP using Q-mixture SM kernel

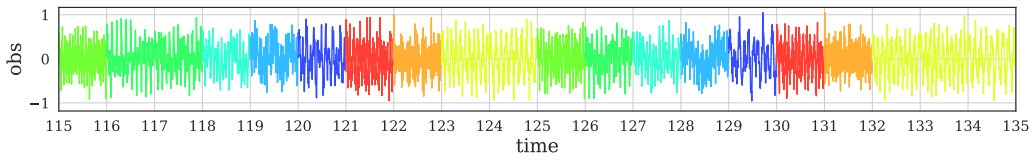
All models are trained by variational inference for full T sequences as explained in section 3.2.

For this experiment, we generate the sequences of time-series observations using Eq. (35) having 150 sequences of time-series. During the experiment, we fix the number of data points in each time-series observation as $N_t = 200$. We limit spectral range as $W_{Hz} = 100$ to make sure that time-series with $N_t = 200$ contains enough information to infer the spectral characteristics of the time-series according to the Nyquist–Shannon sampling theorem (Oppenheim, 1999).

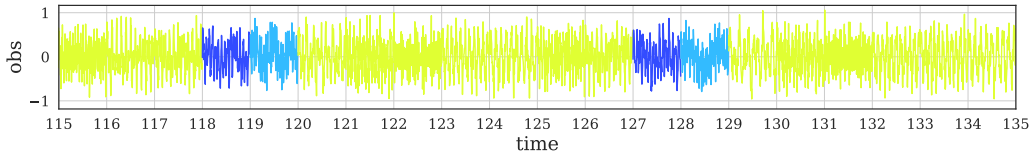
For training the parameters of the model, we use first 100 sequences of time-series, i.e $T = 100$ (training set). We run 30 iterations to update the local variables $q(Z)$ and the variational parameters of $q(\pi)$ and $q(A)$, and kernel hyperparameters for global variables. The kernel hyperparameters are updated by the Adam optimizer with learning rate .005. After training, we evaluate how accurate the trained models estimate the hidden state for the left 50 sequences of time-series observations (test set).

		$W_{Hz} = 100$	
Model	**Inference**	**Accuracy**	**1 Iteration time**
HMM-GPSM (Q=1)	VI (T=100)	.66	98.25
HMM-GPRBF	VI (T=100)	.44	75.05
HMM-GPPER	VI (T=100)	.46	89.12
HMM-GPCombKernel	VI (T=100)	.42	164.23

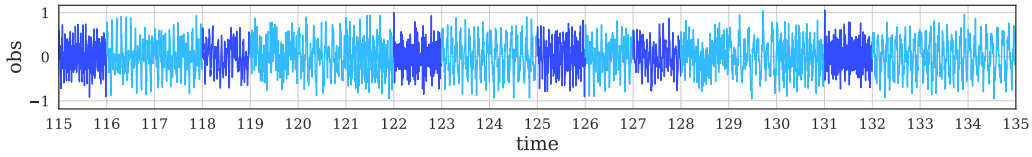
Table 3: Accuracy and single iteration time (seconds) on sinusoidal dataset with $W_{Hz} = 100$; This table compares the statistic of accuracy and single iteration time for HMM-GPSM ($Q = 1$), HMM-GPRBF, HMM-GPPER, and HMM-GPCombKernel. Assuming that each kernel uses the almost same number of hyperparameters (#RBF: 2 , #Periodic: 3), we just use a single mixture SM kernel ($Q = 1$, #SM: 3) to show its expressibility for characterizing the complex time-series observation.



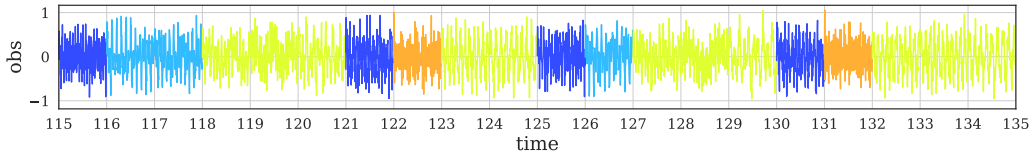
(a) Test data set



(b) HMM-GPRBF trained by VI (T=100) with accuracy 0.44



(c) HMM-GPPER trained by VI (T=100) with accuracy 0.46



(d) HMM-GPSM (Q=1) trained by VI (T=100) with accuracy 0.66

Figure 6: Test set clustering results of HMM-GPSM, HMM-GPPER, and HMM-GPRBF for $W_{Hz} = 100$. We display the estimated result $\{\hat{z}_t\}_{t=115}^{134}$ on $[115, 135]$ out of $[100, 150]$ because of limited space. We see that SM kernel allows the model to estimate hidden states more flexibly and accurately over RBF and Periodic kernel; (d) SM kernel distinguishes the testset into 4 types of signal. (b) RBF and (c) Periodic kernel distinguish the testset into 3 types and 2 types of signal.