Variational inference for state space models: theoretical guarantees, practical implementation and online learning

Sylvain Le Corff
based on joint works with
Hermanni Hälvä, Luc Lehéricy, Élisabeth Gassiat, Aapo Hyvärinen
Pierre Gloaguen, Jimmy Olsson, Mathis Chagneux

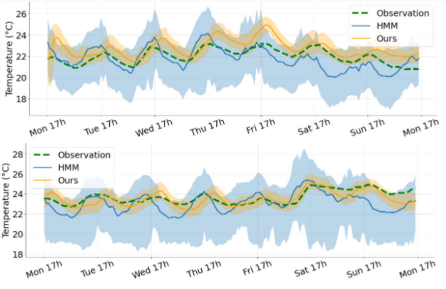
LPSM, Sorbonne Université

Motivations

Overview - learn *disentangled* representations

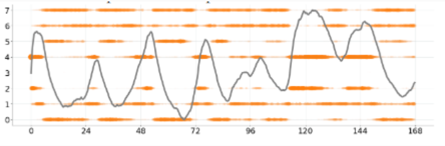
- Common assumption in unsupervised representation learning: **low-dimensional latent variables generate observed data**.
- Knowledge of ***true* latent variables** useful in many tasks: classification, transfer learning, causal inference etc.
- Problem: **models used usually unidentifiable** (e.g. β-VAE), thus we cannot recover *true* data generating features.
- Contributions: **general identifiable** framework for principled disentanglement. Deep leargning architectures for **structured VAE**. **Some theoretical guarantees** for VI for state spaces.

Applications - building energy management - Cohen, M. et al., 2021



$\rightharpoondown$ $(x_k)_{k\geqslant 0}$: observations to be predicted - indoor temperatures, consumptions, humidity levels in large buildings.

$\rightharpoondown$ Latent states $(s_k)_{k\geqslant 0}$: used to identify random sollicitations (meteorological) and usages.



$\rightharpoondown$ **Efficient training algorithms** for overly large deep learning models. **Identification of the latent states**.

Identifiability from dependent data

Deconvolution

The observation $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{Z} + \varepsilon \, ,$$

$\mathbf{Z}$ is the signal and ε is the noise, $\mathbf{Z}$ and ε are independent random variables.

Goal

Learn the distribution of $\mathbf{Z}$ and of ε using independent observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ only.

Constraints

"No assumptions" on the distribution of the noise ε.

We do not assume that some samples with the same distribution as ε are available.

First basic question: identifiability

Is the distribution of $\mathbf{Z}$ uniquely determined by the distribution of $\mathbf{X}$? That is:

Can $\mathbf{Z} + \varepsilon$ have the same distribution of $\mathbf{Z}' + \varepsilon'$ with $\mathbf{Z}'$ having a different distribution than $\mathbf{Z}$?

What assumptions to get identifiability (up to translation) ?

Good news: no assumption on the noise and weak structure assumptions on the signal allow identifiability

- Multidimensional observations: $\mathbf{X}$, $\mathbf{Z}$, ε are in $\mathbb{R}^d$, $d \geq 2$
- No distributional assumption on the noise, except that it has independent components
- The distribution of the signal has not too heavy tails
- Some dependency assumption on the components of the signal

Identifiability theorem

- With $d_1 \geqslant 1$, $d_2 \geqslant 1$ $(d_1 + d_2 = d)$:

$$\mathbf{X} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} + \begin{pmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \end{pmatrix} = \mathbf{Z} + \varepsilon \,.$$

- $\varepsilon^{(1)}$ is independent of $\varepsilon^{(2)}$.

$\mathbb{P}_{R,Q}$ is the distribution of $\mathbf{X}$ when $\mathbf{Z}$ has distribution R and for $i \in \{1, 2\}$, $\varepsilon^{(i)}$ has distribution $Q^{(i)}$, with $Q = Q^{(1)} \otimes Q^{(2)}$.

- "Dependency assumption" on $X^{(1)}$ and $X^{(2)}$ (HD).

- Tail assumption on R (H(ρ)).

Theorem

Assume that R and $\tilde{R}$ are probability distributions on $\mathbb{R}^d$ which satisfy assumption H(ρ) for some $\rho < 2$ and which satisfy HD.
Then, $\mathbb{P}_{R,Q} = \mathbb{P}_{\tilde{R},\widetilde{Q}}$ implies that $R = \tilde{R}$ and $Q = \widetilde{Q}$ up to translation.

Application to nonlinear ICA

Structured Nonlinear ICA & Examples

Observations: $(\boldsymbol{x}_t)_{t\in\mathbb{T}} = ((x_t^{(1)}, \ldots, x_t^{(M)}))_{t\geqslant 0}$.

Independence of latent components:

$$p(\boldsymbol{s}_{t_1}, \ldots, \boldsymbol{s}_{t_m}) = \prod_{i=1}^{N} p(s_{t_1}^{(i)}, \ldots, s_{t_m}^{(i)}) \,.$$

Nonlinear observation model:

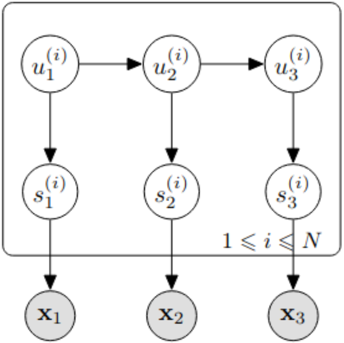
$$\boldsymbol{x}_t = \mathbf{f}(\boldsymbol{s}_t) + \varepsilon_t \,,$$

where $(\varepsilon_t)_{t\geqslant 1}$ are i.i.d with **unknown distribution**; $\boldsymbol{f} : \mathbb{R}^N \to \mathbb{R}^M$ is injective.

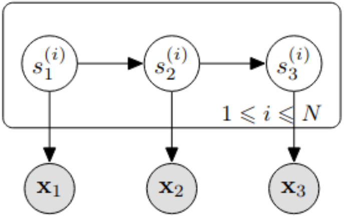
(Hyvarinen, A. and Pajunen, P., 1999, Neural Networks):
"noise free" nonlinear ICA not identifiable i.e. infinitely many decompositions of $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s})$ into independent components.

(Hyvarinen, A., Sasaki, H., and Turner R., 2019, AISTATS):
independent components dependent on some additional auxiliary variable $\boldsymbol{u}$, while being conditionally mutually independent.

Previous models can be reformulated to fit within our framework.

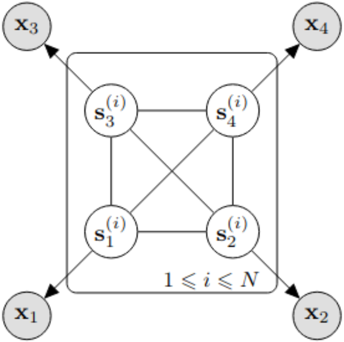


(a) HMM modulated components c.f. (Hälvä and Hyvärinen, 2020)

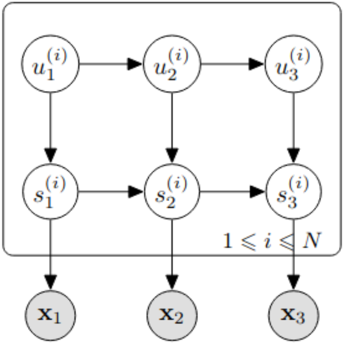


(b) Temporal dependencies c.f. (Hyvärinen and Morioka, 2017)

As well as flexible new models.



(c) New: Spatial process on a graph (with latent states u_t integrated out)



(d) New: Δ-SNICA , a linear switching dynamics model for components

Step 1: identification of noise free distribution

$\rightharpoondown$ **Identify noise-free distribution of $\boldsymbol{z}_t = \boldsymbol{f}(\boldsymbol{s}_t)$ from $\boldsymbol{x}_t = \boldsymbol{z}_t + \boldsymbol{\varepsilon}_t$.**

$\rightharpoondown$ **Assumptions**

- (A1) Tails of $\boldsymbol{z}_t$ "not much" heavier than Gaussian.
  For some $\rho < 3$, for all $\lambda \in \mathbb{R}^M$, $\mathbb{E}[\exp(\lambda^\top \boldsymbol{z}_t)] \leqslant A \exp(B\|\lambda\|^\rho)$.
- (A2) Non-degeneracy assumption.
  The random variables $(\boldsymbol{z}_t)_{t \geqslant 1}$ are dependent.
- (A3) $\boldsymbol{z}_t$ has no Gaussian component.

If (A1), (A2) and (A3) hold for some $(t_1, t_2) \in \mathbb{T}^2$. Then, for all $m \geqslant 2$, the law of **$(\boldsymbol{z}_{t_1}, \dots, \boldsymbol{z}_{t_m})$ and the law of ε_{t_1} can be recovered** up to translation **from the law of $(\boldsymbol{x}_{t_1}, \dots, \boldsymbol{x}_{t_m})$**.

Step 2: identification of the mixing function

→ Noise ε can have arbitrary and unknown distribution! Similar as (Gassiat É., Le Corff, S. and Lehéricy, L., 2020, JMLR) and (Gassiat É., Le Corff, S. and Lehéricy, L., 2022, AoS).

→ **Identify f from the distribution of $(z_{t_1}, \dots, z_{t_m})$.**

→ Under additional technical assumptions, **f can be recovered** up to permutation and component-wise transformations **from the law of $(z_{t_1}, \dots, z_{t_m})$**.

Practical implementation - Δ-SNICA model

Labels: $(\boldsymbol{u}_t)_{t \geqslant 1}$ discrete Markov chain in $\{1, \ldots, K\}$.

Regime switching: For all $1 \leqslant i \leqslant N$, $t \geqslant 2$, $\boldsymbol{y}_t^i = B^i_{u_t^i} \boldsymbol{y}^i_{t-1} + b_{u_t^i} + \varepsilon^i_{u_t^i}$.

Target signals: The independent components are $\boldsymbol{s}_t^i = \boldsymbol{y}^i_{t,1}$.

Observation model: The observations are $\boldsymbol{x}_t = \boldsymbol{f}_\theta(\boldsymbol{s}_t) + \eta_t$, with $(\eta_t)_{t \geqslant 1}$ i.i.d. and Gaussian.

Parameters: Law of the discrete chain, parameters of the linear and Gaussian state space model, parameters of f_θ (typically a Feed Forward Neural Network).

The loglikelihood **cannot be computed**, in this work we use a **variational formulation**.

Variational estimation

In practice the model is often estimated by **maximizing the ELBO**:

$$\mathcal{L}(\theta, \varphi, \boldsymbol{x}_{1:t}) = \mathbb{E}_{q_{\varphi,0:t}}\left[\log \frac{p_\theta(\boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t})}{q_{\varphi,0:t}(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})}\right]$$

where $q_{\varphi,0:t}(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})$ is the **variational distribution**.

Traditional assumption on the variational family

$$q_{\varphi,0:t}(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t}) = \prod_{s=1}^{t} q_{\varphi,s}(\boldsymbol{z}_s|\boldsymbol{x}_{1:t})\,.$$

→ No theoretical results and does not fit classical posterior distributions (for instance in HMMs).

New framework: backward decomposition

$$q_{\varphi,0:t}(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t}) = q_{\varphi,t}(\boldsymbol{z}_t|\boldsymbol{x}_{1:t}) \prod_{s=2}^{t} q_{\varphi,s-1|s}(\boldsymbol{z}_{s-1}|\boldsymbol{z}_s, \boldsymbol{x}_{1:t})\,.$$

→ Some theoretical guarantees and well designed for online learning.

A few theoretical results for reconstruction

Setting

State space models

$$\underbrace{\phi_{0:t}^{\theta}}_{Z_{0:t} \text{ given } X_{0:t}} h = \mathbb{E}_{\theta}\left[h(Z_{0:t}) | X_{0:t}\right]$$

- $Z_{0:t}$ is a Markov chain with transition density m_θ.
- Conditionally on $Z_{0:t}$, the observations are independent with emission densities $g_\theta(Z_t, \cdot)$.

Additive state functionals

$$h_{0:t} : z_{0:t} \mapsto \sum_{s=1}^{t} \tilde{h}_s(z_{s-1}, z_s)$$

$\rightsquigarrow$ $\phi_{0:t}^{\theta} h_{0:t}$ *crucial in both inference and parameter learning.*

Goal - bias control

Theoretically validate backward variational smoothing as a valid approximation.

- Variational inference **is not consistent**.
- Bias depends on implementation / optimization.

$\rightsquigarrow$ Ensure that the bias is controlled w.r.t time.

Quantities of interest: $\phi^{\theta}_{0:t} h_{0:t} = \mathbb{E}_{\theta}\left[h_{0:t}(Z_{0:t}) | X_{0:t}\right]$

$h_{0:t}$ additive state functional.

$$|q_{\varphi,0:t} h_{0:t} - \phi^{\theta}_{0:t} h_{0:t}| \leq ?$$

$\rightsquigarrow$ Marginal smoothing as a byproduct.

An insightful result

Assumptions

- $\sigma_- \leq \ell_s^\theta(x_{s-1}, x_s) \leq \sigma_+$ and $\sigma_- \leq q_{s-1|s}^\lambda(x_{s-1}, x_s) \leq \sigma_+$
- $\left\| q_{\varphi,t} - \phi_t^\theta \right\|_{\mathrm{tv}} \leq \varepsilon$.
- $\left\| q_{\varphi,s-1|s}(x_s, \cdot) - b_{s-1|s}^\theta(x_s, \cdot) \right\|_{\mathrm{tv}} \leq \varepsilon$ for all $s < t,\ x_s \in \mathsf{X}$.

Additive bound

$$\left| q_{\varphi,0:t} h_{0:t} - \phi_{0:t}^\theta h_{0:t} \right| \leq c t \varepsilon$$

Perspectives

Quantitative bounds without strong mixing ?
Does minimizing the ELBO ensure that the true and variational kernels are close ?

To obtain excess risk bound

Assumptions

- $\sigma_- \leq \ell_s^\theta(x_{s-1}, x_s) \leq \sigma_+$ and $\sigma_- \leq q_{s-1|s}^\lambda(x_{s-1}, x_s) \leq \sigma_+$
- $\mathrm{KL}(q_{\varphi,t}, \phi_t^\theta) \leq \varepsilon$.
- $\mathrm{KL}(q_{\varphi,s-1|s}(x_s, \cdot), b_{s-1|s}^\theta(x_s, \cdot)) \leq \varepsilon$ for all $s < t$.
- Additional moment and Lipschitz assumptions.

There exist constants c_0, c_1, c_2, D such that with probability at least $1 - c_0 \exp(-c_1\{d_* \log n\}^{1\wedge\alpha_*})$, for any $\gamma > 0$,

$$\mathrm{KL}\left(P_{\theta^*} \middle\| P_{\widehat{\theta}_{n,T}}\right) \leq (1+\gamma)(T+1)\epsilon + c_2(1+\gamma^{-1})\frac{Dd_* T^3}{n}\log(d_* n)(\log n)^{1/\alpha_*} \ .$$

There exist constants c_0, c_1, c_2, D such that with probability at least $1 - c_0 \exp(-c_1\{d_* \log n\}^{1\wedge\alpha_*})$, for any $\gamma > 0$,

$$\mathrm{KL}\left(P_{\theta^*} \middle\| P_{\widehat{\theta}_{n,T}}\right) \leq (1+\gamma)(T+1)\epsilon + c_2(1+\gamma^{-1})\frac{Dd_* T^3}{n}\log(d_* n)(\log n)^{1/\alpha_*} .$$

Perspectives

Improving the dependency with respect to T ?
Specific results (constants) for specific deep architectures ?

To obtain excess risk bound

Write

$$\widehat{P}_n(\mathrm{d}x_{0:t}) = \int \left(\frac{1}{n} \sum_{i=1}^{n} q_\varphi(z_{0:t}|x_{0:t}^i) \right) p_\theta(x_{0:t}|z_{0:t}) \mathrm{d}z_{0:t}$$

There exist constants c_0, c_1, c_2, D such that with probability at least $1 - c_0 \exp(-c_1 \{d_* \log n\}^{1\wedge\alpha_*})$, for any $\gamma > 0$,

$$\mathrm{KL}\left(P_{\theta^*} \middle\| \widehat{P}_n\right) \leq (1+\gamma)(T+1)\epsilon + c_2(1+\gamma^{-1}) \frac{Dd_* T^3}{n} \log(d_* n)(\log n)^{1/\alpha_*} .$$

Perspectives

Improving the dependency with respect to T ?
Specific results (constants) for specific deep architectures ?

Deep learning-based implementations

Practical implementation - Δ-SNICA model

Labels: $(\boldsymbol{u}_t)_{t\geqslant 1}$ discrete Markov chain in $\{1,\dots,K\}$.

Regime switching: For all $1\leqslant i\leqslant N$, $t\geqslant 2$, $\boldsymbol{y}_t^i = B^i_{u_t^i}\boldsymbol{y}^i_{t-1} + b_{u_t^i} + \varepsilon^i_{u_t^i}$.

Target signals: The independent components are $\boldsymbol{s}_t^i = \boldsymbol{y}^i_{t,1}$.

Observation model: The observations are $\boldsymbol{x}_t = \boldsymbol{f}_\theta(\boldsymbol{s}_t) + \eta_t$, with $(\eta_t)_{t\geqslant 1}$ i.i.d. and Gaussian.

Parameters: Law of the discrete chain, parameters of the linear and Gaussian state space model, parameters of f_θ (typically a Feed Forward Neural Network).

The loglikelihood **cannot be computed**, in this work we use a **variational formulation**.

Practical implementation - variational estimation

The model is estimated by **maximizing the ELBO**:

$$\mathcal{L}(\theta, \varphi,)\boldsymbol{x}_{1:t} = \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(\boldsymbol{x}_{1:t}, \boldsymbol{u}_{1:t}, \boldsymbol{y}_{1:t})}{q_\varphi(\boldsymbol{u}_{1:t}, \boldsymbol{y}_{1:t} | \boldsymbol{x}_{1:t})}\right]$$

where $q_\varphi(\boldsymbol{u}_{1:t}, \boldsymbol{y}_{1:t} | \boldsymbol{x}_{1:t})$ is the **variational distribution**.

Assumption (I) on the variational family:

$$q_\varphi(\boldsymbol{u}_{1:t}, \boldsymbol{y}_{1:t} | \boldsymbol{x}_{1:t}) = q_\varphi(\boldsymbol{u}_{1:t} | \boldsymbol{x}_{1:t}) q_\varphi(\boldsymbol{y}_{1:t} | \boldsymbol{x}_{1:t}) \,.$$

Using the assumption on the model all terms of the ELBO can be computed except $\mathbb{E}_{q_\varphi}[\sum_{s=1}^{t} \log p_\theta(\boldsymbol{x}_t | \boldsymbol{s}_t)]$ which is approximated using other neural nets.

$\rightarrow$ Allows very fast variational learning but no theoretical guarantees for such approaches.

Practical implementation - variational estimation

The model is estimated by **maximizing the ELBO**:

$$\mathcal{L}(\theta, \varphi, \boldsymbol{x}_{1:t}) = \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(\boldsymbol{x}_{1:t}, \boldsymbol{u}_{1:t}, \boldsymbol{y}_{1:t})}{q_\varphi(\boldsymbol{u}_{1:t}, \boldsymbol{y}_{1:t} | \boldsymbol{x}_{1:t})}\right]$$

where $q_\varphi(\boldsymbol{u}_{1:t}, \boldsymbol{y}_{1:t} | \boldsymbol{x}_{1:t})$ is the **variational distribution**.

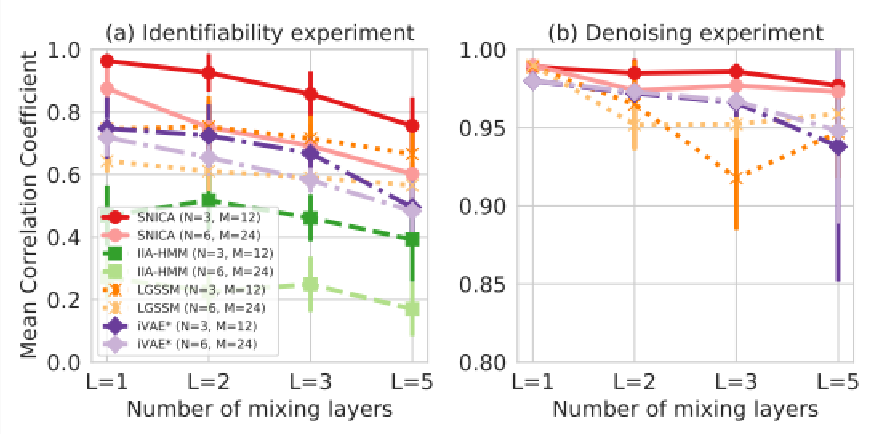
Assumption (II) on the variational family:

$$q_\varphi(\boldsymbol{u}_{1:t}, \boldsymbol{y}_{1:t} | \boldsymbol{x}_{1:t}) = q_\varphi(\boldsymbol{u}_t, \boldsymbol{y}_t | \boldsymbol{x}_{1:t}^{1:N}) \prod_{s=1}^{t-1} q_\varphi(\boldsymbol{u}_s, \boldsymbol{y}_s | \boldsymbol{u}_{s+1}, \boldsymbol{y}_{s+1}, \boldsymbol{x}_{1:s}) \,.$$

$\rightharpoondown$ Allows online learning and first theoretical guarantees for such approaches.

Experiments (I.1)

- 100k-long time series sampled from the model, $K = 2$.
- Observed data of dimension $M \in \{12, 24\}$ - number of independent components, $N \in \{3, 6\}$.
- We considered four levels of mixing of increasing complexity by randomly initialized MLPs of the following number of layers: 1 (linear ICA), 2, 3, and 5.
- Simulated data: **Measure identifiability** – correlation between estimated and true independent components.

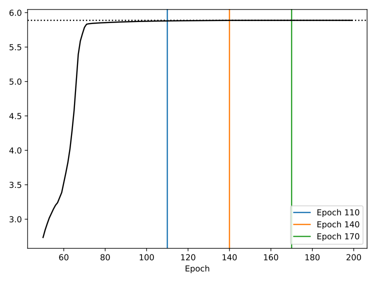


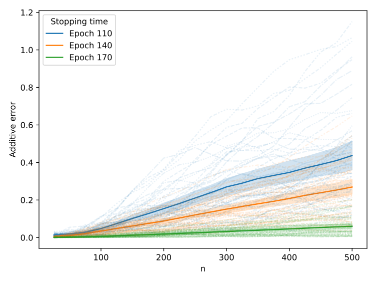
(a) Identifiability experiment
Mean Correlation Coefficient
1.0
0.8
0.6
0.4
0.2
0.0
L=1
L=2
L=3
L=5
Number of mixing layers
SNICA (N=3, M=12)
SNICA (N=6, M=24)
IIA-HMM (N=3, M=12)
IIA-HMM (N=6, M=24)
LGSSM (N=3, M=12)
LGSSM (N=6, M=24)
iVAE* (N=3, M=12)
iVAE* (N=6, M=24)
(b) Denoising experiment
1.00
0.95
0.90
0.85
0.80
L=1
L=2
L=3
L=5
Number of mixing layers

Experiments (II.1)

- Use of backward variational law to illustrate theoretical results - errors grow linearly with the number of obwervations for additive functionals.
- True observation model given by a Gaussian law with mean $h_\theta(s_t)$ and variance R.
- Hidden signals given by a linear and Gaussian state-space.
- Variational backward kernel given by a Gaussian law with DNN to encode means and variance.

Experiments (II.2)



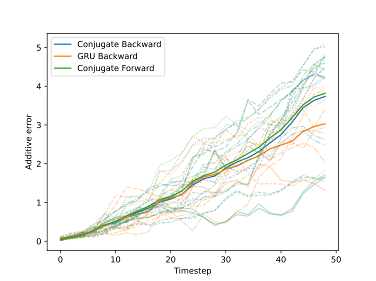


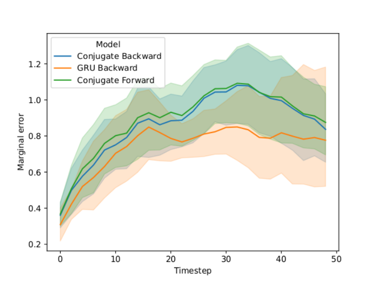
(Left) Model trained and stop after different number of epochs.

(Right) State estimation error for an additive functional for each variational model.

Experiments (III)

For all k, conditionally on z_{k-1}, z_k is Gaussian with mean $z_{k-1} + \delta[\gamma W \tanh(z_{k-1}) - z_{k-1}]/\tau$ and variance Q and the emission density is a Student-t distribution with mean z_k, ν degrees of freedom and scale R.





(Left) Smoothing errors at each time step (10 independent runs).

(Right) Marginal smoothing errors at each time step (10 independent runs).

Challenges

- Design new methodologies for more general variational families (non Gaussian noise, etc.).
- Large scale online learning.
- Theoretical guarantees with weaker assumptions (forgetting, consistency).
- Theoretical guarantees for online learning.