Clinical Time Series Imputation using Conditional Information Bottleneck

MinGyu Choi, Changhee Lee

DEEP GENERATIVE MODELS FOR HEALTH @ NeurIPS 2023

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

TL;DR: Conditional information bottleneck (CIB) enables mitigating strict regularization of information bottleneck while preserving temporal dynamics.

Clinical time series imputation presents a significant challenge because it requires capturing the underlying temporal dynamics from partially observed time series data input. However, direct application of ***information bottleneck* (IB)** framework to time series data without considering temporal context can lead to a substantial loss of temporal dependencies, which, in turn, can degrade the overall imputation performance and further clinical decisions.

To address such a challenge, we propose a novel ***conditional information bottleneck* (CIB)** approach for time series imputation. Variational decomposition of CIB motivates us to develop a novel deep learning method that can approximately achieve the proposed CIB objective for time series imputation as a combination of evidence lower bound and novel temporal kernel-enhanced contrastive optimization.

Information Bottleneck (IB) on Imputation

Definition 1. ***(Imputation)*** *Let $\mathbf{X}^o$ and $\mathbf{X}^m$ be random variables for the partially observed features and missing features of $\mathbf{X}$, respectively, such that $\mathbf{X} = \mathbf{X}^o \cup \mathbf{X}^m$. Then, we define imputation as an unsupervised IB as follows:*

$$\min_{\phi,\theta} \; I_\phi(\mathbf{Z}; \mathbf{X}^o) - \beta I_\theta(\mathbf{X}; \mathbf{Z}) \tag{1}$$

where $\beta \in \Re_>$ is a Lagrangian multiplier, and ϕ and θ correspond to learnable parameters that define probabilistic mappings $q_\phi(\mathbf{Z}|\mathbf{X}^o)$ and $q_\theta(\mathbf{X}|\mathbf{Z})$, respectively.

Goal: finding the distribution of latent representation **Z** and the corresponding parameters that preserves the core information for accurately reconstructing the (complete) original input **X** while suppressing redundant information from its incomplete observation, $\mathbf{X}^o$

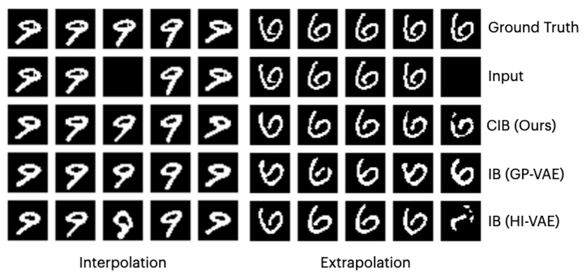

Figure 1B. Motivating experimental results on interpolation (left) and extrapolation (right). Because features in a single time step are completely missing, a model must collect information from other time steps. The conventional IB approach (HI-VAE) shows deteriorating performance in both cases. Another IB approach (GP-VAE) using a Gaussian process prior demonstrates enhanced performance for interpolation but often significantly loses time series characteristics for extrapolation (i.e., the writing style is corrupted). The CIB approach (Ours) exhibits improved imputation performance for both cases.

Unconditional regularization may lead to a significant loss of information regarding the temporal context. Therefore, apply regularization conditioned on remaining time steps, conditional information bottleneck (CIB).

Conditional Information Bottleneck (CIB)

Definition 2. ***(Time Series Imputation)*** *Let $\mathbf{X}_t^o$ and $\mathbf{X}_t^m$ be random variables for the partially observed features and missing features of $\mathbf{X}_t$ at time step t. Then, given the observed time series input $\mathbf{X}_{1:T}^o$, we define time series imputation at time step t as an unsupervised CIB as follows:*

$$\min_{\phi,\theta} \; \underbrace{I_\phi(\mathbf{Z}_t; \mathbf{X}_t^o | \mathbf{X}_{\setminus t}^o)}_{\textit{Conditional Regularization}} - \beta \underbrace{I_\theta(\mathbf{X}_t; \mathbf{Z}_t)}_{\textit{Reconstruction}} \tag{2}$$

where $\mathbf{X}_{\setminus t}^o$ represents the random variables for the remaining input observations, excluding $\mathbf{X}_t^o$.

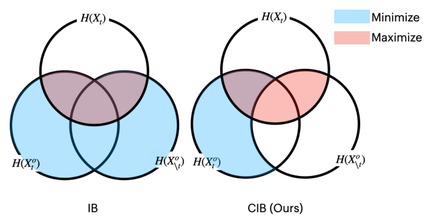

Figure 1A. Conceptual illustration of the IB and CIB principles.

Variational Decomposition of CIB

Maximizing Reconstruction $\min_{\phi,\theta} -I(\mathbf{X}_t; \mathbf{Z}_t)$

$$I_\theta(\mathbf{X}_t; \mathbf{Z}_t) \geq \mathbb{E}_{\mathbf{x}_{1:T}^o \sim p_{\text{data}}} \left[\mathbb{E}_{\mathbf{z}_t \sim q_\phi(\mathbf{z}_t | \mathbf{x}_{1:T}^o)} \left[\log p_\theta(\mathbf{x}_t | \mathbf{z}_t) \right] \right] \stackrel{\text{def}}{=} -\mathcal{L}_{\phi,\theta}^1 \tag{3}$$

(Voloshynovskiy et al., 2019)

Minimizing Conditional Regularization $\min_{\phi,\theta} I_\phi(\mathbf{Z}_t; \mathbf{X}_t^o | \mathbf{X}_{\setminus t}^o)$

Chain rule of mutual information gives:

$$\min_{\phi,\theta} I(\mathbf{Z}_t; \mathbf{X}_t^o | \mathbf{X}_{\setminus t}^o) = \min_{\phi,\theta} I(\mathbf{Z}_t; \mathbf{X}_{1:T}^o) - I(\mathbf{Z}_t; \mathbf{X}_{\setminus t}^o). \tag{4}$$

1. Conciseness term:

$$I(\mathbf{Z}_t; \mathbf{X}_{1:T}^o) \leq \mathbb{E}_{\mathbf{x}_{1:T}^o \sim p_{\text{data}}} \left[D_{\text{KL}}(q_\phi(\mathbf{z}_t | \mathbf{x}_{1:T}^o) || p(\mathbf{z}_t)) \right] \stackrel{\text{def}}{=} \mathcal{L}_\phi^2 \tag{5}$$

2. Temporal dynamics term:

$$I(\mathbf{Z}_t; \mathbf{X}_{\setminus t}^o) \geq \mathbb{E}_{\mathbf{x}_{1:T}^o \sim p_{\text{data}}} \left[\log \left(\frac{\sum_{t' \in \{1,\dots,T\} \setminus t} \exp\left(\mathbf{z}_t{}^T \tilde{\mathbf{z}}_{t'} / \tau \right)}{\sum_{\mathbf{x}_{1:T}^- \in \mathcal{X}_{1:T}^-} \sum_{\substack{t' \in \{1,\dots,T\}, \\ \mathbf{z}_{t'}^- \sim q_\phi(\mathbf{z}_{t'}^- | \mathbf{x}_{1:T}^-)}} \exp\left(\mathbf{z}_t{}^T \mathbf{z}_{t'}^- / \tau \right)} \right) \right] \stackrel{\text{def}}{=} -\mathcal{L}_\phi^3 \tag{6}$$

(A. v. d. Oord et al., 2018)

Introducing inductive bias on temporal dynamics:

Temporal kernels with Conditional alignment

$$I(\mathbf{Z}_t; \mathbf{X}_{\setminus t}^o) \geq \mathbb{E}_{\mathbf{x}_{1:T}^o \sim p_{\text{data}}} \left[\log \left(\frac{\sum_{t' \in \{1,\dots,T\} \setminus t} c_{t,t'} \; \exp\left(\mathbf{z}_t{}^T \tilde{\mathbf{z}}_{t'} / \tau \right)}{\sum_{\mathbf{x}_{1:T}^- \in \mathcal{X}_{1:T}^-} \sum_{\substack{t' \in \{1,\dots,T\}, \\ \mathbf{z}_{t'}^- \sim q_\phi(\mathbf{z}_{t'}^- | \mathbf{x}_{1:T}^-)}} \exp\left(\mathbf{z}_t{}^T \mathbf{z}_{t'}^- / \tau \right)} \right) \right] \stackrel{\text{def}}{=} -\mathcal{L}_\phi^{3'} \tag{7}$$

(Dufumier et al., 2019)

$$c_{\text{cauchy}}(\tau, \tau') = \sigma^2 \left(1 + (\tau - \tau')^2 / l^2 \right)^{-1}$$

Experiments

Image Sequence Dataset

Methods	**HealingMNIST** (*missing with MNAR pattern*)			**RotatedMNIST** (*interpolation & extrapolation*)	
	NLL(↓)	MSE(↓)	AUROC(↑)	NLL(↓)	MSE(↓)
No Imp.	-	0.293 ± 0.000	0.920 ± 0.000	-	0.133 ± 0.000
Mean Imp.	-	0.168 ± 0.000	0.938 ± 0.000	-	0.085 ± 0.000
Forward Imp.	-	0.177 ± 0.000	0.946 ± 0.000	-	0.080 ± 0.000
VAE	0.480 ± 0.002	0.232 ± 0.000	0.922 ± 0.000	1.773 ± 0.127	0.133 ± 0.000
HI-VAE	0.290 ± 0.001	0.134 ± 0.003	0.962 ± 0.001	0.207 ± 0.007	0.087 ± 0.001
GP-VAE	0.261 ± 0.001	0.114 ± 0.002	0.960 ± 0.002	0.190 ± 0.001	0.080 ± 0.004
Ours(Uniform)	0.204 ± 0.002	0.090 ± 0.001	$\mathbf{0.967} \pm 0.001$	$\mathbf{0.184} \pm 0.001$	0.077 ± 0.001
Ours(Cauchy)	$\mathbf{0.202} \pm 0.004$	$\mathbf{0.088} \pm 0.002$	$\mathbf{0.967} \pm 0.000$	$\mathbf{0.184} \pm 0.001$	$\mathbf{0.076} \pm 0.002$

Table 1. Imputation and prediction performance on Image sequence dataset.

Robustness Analysis

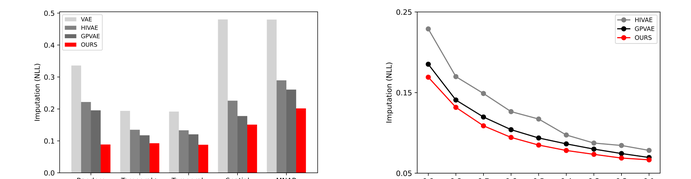

Figure 2. Robustness on missing patterns (Left) and missing ratios (Right). *Spatial* (i.e., neighboring pixels have correlated missing probabilities), and *Temporal+/-* (i.e., positive/negative temporal correlation). Missing ratio experiments performed with random missing pattern.

Physionet2012

Methods	**Physionet2012** (*mortality prediction*)		
	NLL(↓)	MSE(↓)	AUROC(↑)
No Imp.	-	0.962 ± 0.000	0.692 ± 0.000
Mean Imp.	-	0.511 ± 0.000	0.703 ± 0.000
Forward Imp.	-	0.613 ± 0.000	0.710 ± 0.000
BRITS	-	0.529 ± 0.004	0.700 ± 0.005
SAITS	-	$\mathbf{0.501} \pm 0.024$	0.713 ± 0.007
VAE	1.400 ± 0.000	0.962 ± 0.000	0.691 ± 0.001
HI-VAE	1.345 ± 0.009	0.852 ± 0.018	0.696 ± 0.004
GP-VAE	1.227 ± 0.007	0.616 ± 0.013	0.730 ± 0.006
Ours(Uniform)	1.183 ± 0.007	0.528 ± 0.014	$\mathbf{0.744} \pm 0.009$
Ours(Cauchy)	$\mathbf{1.179} \pm 0.006$	0.521 ± 0.012	$\mathbf{0.744} \pm 0.009$

Table 2. Imputation and prediction performance on the clinical dataset.

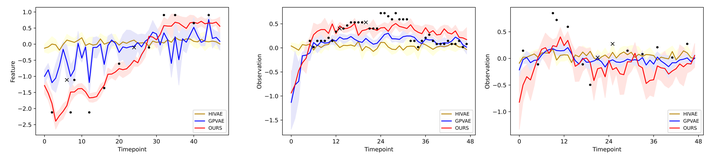

Figure 3. Qualitative results on Physionet2012.

References

B. Dufumier et al., (2021). “Conditional Alignment and Uniformity for Contrastive Learning with Continuous Proxy Labels.” In ArXiv

V. Fortuin et al., (2020). “GP-VAE: Deep Probabilistic Time Series Imputation” In *International conference on artificial intelligence and statistics*.

A. Nazabal et al., (2020). “Handling Incomplete Heterogeneous Data Using VAEs.”. In *Pattern Recognition*.

S. Voloshynovskiy et al., (2019). “Information Bottleneck Through Variational Glasses”. In: *NeurIPS Workshop on Bayesian Deep Learning*.

A. v. d. Oord et al.,(2018). “Representation Learning with Contrastive Predictive Coding.” In: ArXiv.

Openreview