Clinical Time Series Imputation using Conditional Information Bottleneck

MinGyu Choi[*]
MIT

Changhee Lee
Chung-Ang University

Abstract

Clinical time series imputation presents a significant challenge because it requires capturing the underlying temporal dynamics from partially observed time series data input. Among the recent successes of imputation methods based on generative models, the *information bottleneck* (IB) framework offers a well-suited theoretical foundation for multiple imputations, allowing us to account for the uncertainty associated with the imputed values. However, direct application of 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. Our experiments, conducted on real-world healthcare dataset and image sequences, demonstrate that our method significantly improves imputation performance, and also enhances prediction performance based on the imputed values.

1 Introduction

Clinical multivariate time series data often includes missing features, with diverse missing ratios and patterns depending on distinct sampling periods or measurement strategies [6]. Since these missing features can significantly impair the medical decisions and comprehension of the temporal dynamics, time series imputation, aiming to reconstruct the missing features, has become a pivotal and pervasive topic in healthcare. What makes time series imputation challenging is that an imputation method must satisfy two requirements: i) it must account for underlying temporal dependencies, and ii) it should allow for *multiple imputations* to facilitate uncertainty quantification for real-world decision-making.

Generative models, particularly variational autoencoders (VAEs)[8], have been employed in the context of multiple imputation tasks due to their capability to generate samples in a probabilistic manner. VAE-based imputation methods primarily focus on defining the evidence lower bound, where the reconstruction error is computed only over the observed part of the incomplete data [10, 16]. These methods can be naturally interpreted under the information bottleneck (IB) principle [18], providing an information-theoretic understanding of what constitutes an imputation-relevant representation. This understanding is based on the fundamental trade-off between maintaining a concise representation (i.e., regularization) and preserving good representation power (i.e., reconstruction).[19]

However, a direct application of the IB principle to time series imputation struggles with capturing the underlying temporal dependencies, as shown in our motivating examples on intrapolation and extrapolation (Figure 1B). In this paper, we theoretically analyze that the overly strict regularization

[*]Work performed while at Chung-Ang University.

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

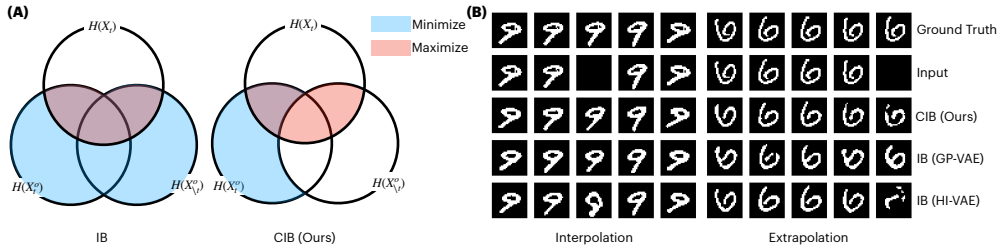

Figure 1: (A) Conceptual illustration of the IB and CIB principles. By conditioning regularization on the remaining input time steps, the latent representation can better preserve the underlying temporal dependency. (B) Motivating experimental results on interpolation (left) and extrapolation (right). 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.

in the conventional IB may force the encoder to rely solely on the particular time point. To overcome such an issue, we propose a novel ***conditional information bottleneck*** (CIB) framework for time series imputation. Our framework adopts the reconstruction-regularization structure of the IB principle while preserving temporal information through conditional regularization, allowing us to circumvent the strict regularization constraints of the conventional IB. Throughout the experiments conducted on healthcare-inspired image sequences and electrical health records (EHR), our proposed method consistently outperforms the state-of-the-art imputation methods with respect to both imputation performance and prediction performance based on the imputed values.

2 Method

2.1 Information Bottleneck Approach to Imputation

We begin with formally define the general imputation task from an information-theoretic perspective. For formal description on information bottleneck on supervised task, please refer to Appendix A.

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 $\beta \in \Re_{\geq}$ 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.

2.2 Conditional Information Bottleneck Approach to Time Series Imputation

We aim to reconstruct the complete time series $\mathbf{x}_{1:T}$ by filling in the missing features from the observed features $\mathbf{x}^o_{1:T}$. Formally, we seek to generate $\mathbf{x}^m_t$ from the conditional distribution $p(\mathbf{X}^m_t|\mathbf{X}^o_{1:T})$.

What makes this problem challenging is that we must account for the underlying temporal dynamics represented by $\mathbf{x}^o_{1:T}$ when imputing missing features $\mathbf{x}^m_t$ for $t \in \{1, \ldots, T\}$. We can straightforwardly apply the unsupervised IB described in (1) by minimizing $I_\phi(\mathbf{Z}_t; \mathbf{X}^o_{1:T}) - \beta I_\theta(\mathbf{X}_t; \mathbf{Z}_t)$ with a comprehensive encoder (e.g., RNN or Transformer). However, enforcing such strict regularization constraints on the encoder may lead to a significant loss of information regarding the temporal context that can be achieved by observations at different time steps. This may cause the imputation of $\mathbf{X}^m_t$ at time step t to heavily rely on the observed features at that particular time point, i.e., $\mathbf{X}^o_t$, rather than being able to learn from temporal dependencies present in other observations, i.e., $\mathbf{X}^o_{\setminus t}$. (Figure 1B)

To tackle this issue, we alleviate the potentially negative consequences of the regularization constraint by directing our attention to the redundant information of the observed input at time step t when it is *conditioned on* its temporal context represented by the remaining observed time series $\mathbf{X}^o_{\setminus t}$. This offers a novel information-theoretic rationale for time series imputation, as defined below:

Definition 2. ***(Time Series Imputation)*** *Let $\mathbf{X}^o_t$ and $\mathbf{X}^m_t$ 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 $\mathbf{X}^o_{1:T}$, we define time series imputation at time step t as an unsupervised CIB as follows:*

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

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

By conditioning on $\mathbf{X}^o_{\setminus t}$, (2) guides us to find latent representations $\mathbf{Z}_t$ and the corresponding inference model parameter ϕ which encompass all retrievable information from the entire observed time series $\mathbf{X}^o_{1:T}$ (***reconstruction***), while discarding information that is redundant for capturing $\mathbf{X}^m_t$ given the available temporal context from the remaining observed time steps $\mathbf{X}^o_{\setminus t}$ (***conditional regularization***). Overall, the proposed objective in (2) enables us to more effectively utilize information from $\mathbf{X}^o_{\setminus t}$ for imputing $\mathbf{X}^m_t$ compared to other IB-related alternatives (see Figure 1A for conceptual illustration).

2.3 Deep Variational Conditional Information Bottleneck on Time Series

In this subsection, we transform (2) into a learnable form by utilizing variational decomposition.

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

Following the derivations introduced in [19], we can find a lower bound of the reconstruction term:

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

Here, we introduce a *feature estimator* $p_\theta(\mathbf{X}_t|\mathbf{Z}_t)$, as a variational approximation of $p(\mathbf{X}_t|\mathbf{Z}_t)$. We model the feature estimator as an isotropic Gaussian, i.e., $p_\theta(\mathbf{X}_t|\mathbf{Z}_t) = \mathcal{N}(\mu_\theta(\mathbf{Z}_t), \text{diag}(\sigma_\theta(\mathbf{Z}_t)))$ where $\mu_\theta(\cdot)$ and $\sigma_\theta(\cdot)$ are implemented by neural networks parameterized by θ.

2.3.2 Minimizing Conditional Regularization: $\min_{\phi,\theta} I_\phi(\mathbf{Z}_t; \mathbf{X}^o_t | \mathbf{X}^o_{\setminus t})$

We employ the chain rule for mutual information on the conditional regularization term as follows:

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

It is worth highlighting that the application of the chain rule decomposes the conditional regularization into two components: (i) minimizing the information between the latent representation $\mathbf{Z}_t$ and the entire observed time series input $\mathbf{X}^o_{1:T}$ that encourages the latent representation to be concise, while (ii) maximizing the information from $\mathbf{X}^o_{\setminus t}$ to capture the underlying temporal dynamics provided by the observations at the remaining time steps. This prevents a significant loss of temporal context in the IB and, in turn, enhances the utilization of temporal dependencies from the remaining time steps.

Minimizing $I(\mathbf{Z}_t; \mathbf{X}^o_{1:T})$ The first term in (4) can be bounded as follows (see Appendix B):

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

where we utilize the unit isotropic Gaussian as the prior distribution, i.e., $p(\mathbf{Z}_t) = \mathcal{N}(\mathbf{0}, I)$. We model the *stochastic encoder* as a multivariate Gaussian distribution defined as $q_\phi(\mathbf{Z}_t|\mathbf{X}^o_{1:T}) = \mathcal{N}(\mu_\phi(\mathbf{X}^o_{1:T}), \text{diag}(\sigma_\phi(\mathbf{X}^o_{1:T})))$, where $\mu_\phi(\cdot)$ and $\sigma_\phi(\cdot)$ are neural networks parameterized by ϕ.

Maximizing $I(\mathbf{Z}_t; \mathbf{X}^o_{\backslash t})$ To bound the second term in (4), we adopt the InfoNCE minimization from the contrastive learning on latent representations that approximately achieves maximizing the corresponding mutual information [11, 17]. We define our novel contrastive learning loss with cosine similarity of latent representations along the time axis as follows (see Appendix C for derivation):

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

where τ is the temperature parameter and $c_{t,t'}$ is a kernel constant (let $c_{t,t'} = 1$ at this stage). Here, $\tilde{\mathbf{z}}_{t'} \sim q_\phi(\tilde{\mathbf{Z}}_{t'} | \mathbf{x}^o_{\backslash t})$ denotes the positive pair obtained by masking the reference time series, such that $\mathbf{x}^o_{\backslash t}$ is created by replacing $\mathbf{x}^o_t$ with zeros from $\mathbf{x}^o_{1:T}$. We regard such positive pairs as augmentations of a given time series since latent representations with missing values at time step t share task-relevant information about the underlying temporal dynamics of a given time series. We denote $\mathcal{X}^-_{1:T}$ a set of negative samples comprising other time series in the same mini-batch, where $\mathbf{x}^-_{1:T}$ indicates an observed time series from $\mathcal{X}^-_{1:T}$. This makes our encoder capture time series-level semantics – such as underlying disease progression patterns that can be distinguished from others – by pushing these samples from the reference. Such an attribution is necessary for reconstructing missing values (and associated downstream tasks in the experiments) specific to the input time series.

Injecting Inductive Bias about Temporal Dynamics The *alignment* of the latent representation [20], attained through contrastive learning without inductive bias, i.e. $c_{t,t'} = 1$ in (6), renders the similarity between latent representations at two adjacent time points indistinguishable from the similarity between those at two distant time points. This phenomenon appears to contradict real-world temporal dynamics, such as gradually deteriorating or periodic behavior of disease progression patterns. To address this, we employ *conditional alignment* [3] that introduces inductive bias about the underlying temporal dynamics with temporal ***Cauchy kernel*** [13] as the following.

$$c_{\text{cauchy}}(\tau, \tau') = \sigma^2 \left(1 + (\tau - \tau')^2 / l^2 \right)^{-1} \tag{7}$$

2.4 Training Objective

We optimize our method based on the following objective by combining all loss functions that allows us to approximately achieve time series imputation defined in (2): $\min_{\phi,\theta} \beta \mathcal{L}^1_{\phi,\theta} + \mathcal{L}^2_\phi + \gamma \mathcal{L}^3_\phi$, where $\gamma \in \mathbb{R}_{\geq 0}$ is a balancing coefficient that trades off the impact of $\mathcal{L}^3_\phi$.

3 Experiments

Evaluation Metrics We evaluate the imputation performance from two perspectives: i) Imputation performance which measures feature-wise (pixel-wise) reconstruction. Specifically, we assess the negative log-likelihood (*NLL*) and mean squared error (*MSE*) of the imputed values on artificially missing features. ii) Prediction performance, which indirectly measures how well the imputed values preserve task-relevant information, which is a crucial aspect of imputation methods in practice. Following [4], we train separate classifiers with imputed values to predict the target labels. Then, we evaluate the area under the receiver operating characteristic (*AUROC*) for classification tasks.

Baseline Models We focus our comparison on VAE-based models since these models can be interpreted under the IB principle as suggested in [19]. Moreover, these multiple imputation methods can provide uncertainty of the imputed values, which is often crucial to support decision-making processes such as clinical interventions in healthcare. Hence, for baseline models, we compare our proposed method with the following: i) **GP-VAE** [4] which utilizes the Gaussian process (GP) prior to model time dependency, ii) **HI-VAE** [10] and iii) **VAE** [8], both of which use an autoencoder architecture and are capable of imputing values at each time step. Note that our model inherits VAE architecture - number of parameters used in encoders and decoders are identical with HI-VAE and VAE; GP-VAE only differs by dimension-wise stochastic encoder. To compare with non-probabilistic model, we also compared with RNN-based **BRITS** [1] and Transformer-based **SAITS** [2]. For a fair comparison, the magnitude of the number of parameters is the same among deep learning methods.

Table 1: Imputation and prediction performance on the image sequence datasets.

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)	$\mathbf{0.202} \pm 0.004$	$\mathbf{0.088} \pm 0.002$	$\mathbf{0.967} \pm 0.000$	$\mathbf{0.184} \pm 0.001$	$\mathbf{0.076} \pm 0.002$

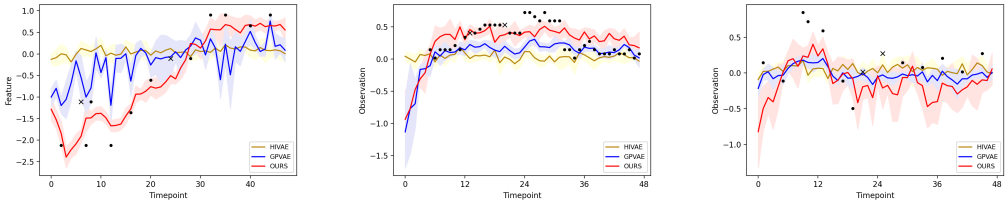

Figure 2: Selected imputation results on Physionet2012.

Imputation on image sequences. To evaluate our model on healthcare-inspired missing scenarios, we assess imputation performance on two MNIST sequence benchmarks. HealingMNIST [9] has approximately 60% of missing pixels under a missing-not-at-random (MNAR) pattern on every time step, where the missing probability of white pixels is twice larger than that of black pixels. Given that the model is not provided with information about the underlying missing mechanism, this task is particularly challenging, yet it mirrors many practical scenarios. For example, in healthcare, patients with depression are more likely to refuse answers about the severity of their condition [5]. RotatedMNIST [12] evaluates performance on interpolation and extrapolation, where all features at an arbitrary time step are completely missing. This makes imputation more challenging since the model must reconstruct all the missing values at a given time step solely based on the temporal dependency. Table 1 demonstrates that our model provides state-of-the-art imputation and prediction performance on both datasets, and Cauchy kernel (7) can further improve the performance.

Imputation for electrical health records. To evaluate on real-world healthcare data, we use *Physionet2012 – Mortality Prediction Challenge* [14], which aims to predict in-hospital mortality of intensive care unit (ICU) patients from 48 hours of records with roughly 80% of missing features. Furthermore, we conduct additional evaluations to assess whether the imputation methods preserve the critical characteristics of a given time series – i.e., whether a patient's status is deteriorating or not – after replacing the missing features with imputed values. Table 2 shows that ours provides imputation performance comparable to the best benchmark while outperforming the VAE-based methods by a great margin. Furthermore, it achieves the best classification performance, successfully capturing information about the temporal dynamics of patients' status. Note that while the SAITS provides the best imputation performance, the imputed values lose the crucial information for discriminating patient's status. We provide qualitative examples comparing with HIVAE and GPVAE, in Figure 2

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

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	$\mathbf{0.744} \pm 0.009$

4 Conclusion

In this paper, we presented a novel information-theoretic approach for clinical time series imputation. CIB addresses the limitation of the IB in capturing underlying temporal dynamics by replacing conventional regularization with conditional regularization. Our empirical results on healthcare-inspired image sequences and electrical health records prove that CIB is effective in practical cases.

References

[1] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li. BRITS: Bidirectional Recurrent Imputation for Time Series. *Advances in neural information processing systems*, 31, 2018.

[2] W. Du, D. Côté, and Y. Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

[3] B. Dufumier, P. Gori, J. Victor, A. Grigis, and E. Duchesnay. Conditional Alignment and Uniformity for Contrastive Learning with Continuous Proxy Labels. *arXiv preprint arXiv:2111.05643*, 2021.

[4] V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt. GP-VAE: Deep Probabilistic Time Series Imputation. In *International conference on artificial intelligence and statistics*, pages 1651–1661. PMLR, 2020.

[5] R. E. Gliklich, N. A. Dreyer, M. B. Leavy, et al. Registries for Evaluating Patient Outcomes: a User's Guide. 2014.

[6] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific data*, 3(1):1–9, 2016.

[7] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised Contrastive Learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[8] D. P. Kingma and M. Welling. Auto-encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 1, 2014.

[9] R. G. Krishnan, U. Shalit, and D. Sontag. Deep Kalman Filters. *arXiv preprint arXiv:1511.05121*, 2015.

[10] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera. Handling Incomplete Heterogeneous Data Using VAEs. *Pattern Recognition*, 107:107501, 2020.

[11] A. v. d. Oord, Y. Li, and O. Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*, 2018.

[12] S. Ramchandran, G. Tikhonov, K. Kujanpää, M. Koskinen, and H. Lähdesmäki. Longitudinal Variational Autoencoder. In *International Conference on Artificial Intelligence and Statistics*, pages 3898–3906. PMLR, 2021.

[13] C. E. Rasmussen. Gaussian Processes in Machine Learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, pages 63–71. Springer, 2004.

[14] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark. Predicting in-hospital Mortality of ICU Patients: The physionet/computing in Cardiology Challenge 2012. In *2012 Computing in Cardiology*, pages 245–248. IEEE, 2012.

[15] K. Sohn. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. *Advances in neural information processing systems*, 29, 2016.

[16] K. Sohn, H. Lee, and X. Yan. Learning Structured Output Representation using Deep Conditional Generative Models. *Advances in neural information processing systems*, 28, 2015.

[17] Y. Tian, D. Krishnan, and P. Isola. Contrastive Multiview Coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[18] N. Tishby and N. Zaslavsky. Deep Learning and the Information Bottleneck Principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015.

[19] S. Voloshynovskiy, M. Kondah, S. Rezaeifar, O. Taran, T. Holotyak, and D. J. Rezende. Information Bottleneck Through Variational Glasses. In *NeurIPS Workshop on Bayesian Deep Learning*, 2019.

[20] T. Wang and P. Isola. Understanding Contrastive Representation Learning Through Alignment and Uniformity on the Hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

A Information Bottleneck on Supervised Tasks.

Let $\mathbf{X}$ and $\mathbf{Y}$ be random variables for the input feature and the target label, respectively. The IB principle aims to find the bottleneck random variable $\mathbf{Z}$ that compresses the information in $\mathbf{X}$ while keeping the information relevant for predicting $\mathbf{Y}$ as the following [18],

$$\min_{\phi,\theta} \; I_\phi(\mathbf{Z};\mathbf{X}) - \beta I_\theta(\mathbf{Y};\mathbf{Z}) \tag{8}$$

where $\beta \in \Re$ is a Lagrangian multiplier that balances the two mutual information terms, and ϕ and θ correspond to learnable parameters that define probabilistic mappings $q_\phi(\mathbf{Z}|\mathbf{X})$ and $q_\theta(\mathbf{Y}|\mathbf{Z})$, respectively. The core motivation of (8) is to find the optimal distribution of latent representation $\mathbf{Z}$ and the corresponding inference model parameters ϕ that removes label-irrelevant information from $\mathbf{X}$ while preserving the information about the class label $\mathbf{Y}$.

B Variational Approximation of First Term in Conditional Regularization.

$$\begin{aligned}
I(\mathbf{Z}_t;\mathbf{X}^o_{1:T}) &= \mathbb{E}_{q_\phi(\mathbf{z}_t,\mathbf{x}^o_{1:T})}\left[\log \frac{q_\phi(\mathbf{x}^o_{1:T},\mathbf{z}_t)}{q_\phi(\mathbf{z}_t)p_{\text{data}}(\mathbf{x}^o_{1:T})}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}_t,\mathbf{x}^o_{1:T})}\left[\log \frac{q_\phi(\mathbf{z}_t|\mathbf{x}^o_{1:T})}{q_\phi(\mathbf{z}_t)}\frac{p(\mathbf{z}_t)}{p(\mathbf{z}_t)}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}_t,\mathbf{x}^o_{1:T})}\left[\log \frac{q_\phi(\mathbf{z}_t|\mathbf{x}^o_{1:T})}{p(\mathbf{z}_t)}\right] + \mathbb{E}_{q_\phi(\mathbf{z}_t,\mathbf{x}^o_{1:T})}\left[\log \frac{p(\mathbf{z}_t)}{q_\phi(\mathbf{z}_t)}\right] \\
&= \mathbb{E}_{p_{\text{data}}(\mathbf{x}^o_{1:T})}\left[D_{\text{KL}}(q_\phi(\mathbf{z}_t|\mathbf{x}^o_{1:T})||p(\mathbf{z}_t))\right] - D_{\text{KL}}(q_\phi(\mathbf{z}_t)||p(\mathbf{z}_t)) \\
&\leq \mathbb{E}_{p_{\text{data}}(\mathbf{x}^o_{1:T})}\left[D_{\text{KL}}(q_\phi(\mathbf{z}_t|\mathbf{x}^o_{1:T})||p(\mathbf{z}_t))\right]
\end{aligned} \tag{9}$$

The last inequality holds because of the non-negativity of KL-divergence.

C Contrastive Approximation of Second Term in Conditional Regularization.

In Section 2.3.2, we optimize $I(\mathbf{Z}_t; \mathbf{X}^o_{\backslash t})$ by approximate the mutual information into a contrastive form, which is similar to [11, 17].

$$
\begin{aligned}
I(\mathbf{Z}_t; \mathbf{X}^o_{\backslash t}) &= -\mathbb{E}_X \log \left[\frac{p(\mathbf{X}^o_{\backslash t})}{p(\mathbf{X}^o_{\backslash t}|\mathbf{Z}_t)} \right] \\
&= -\mathbb{E}_X \log \left[\frac{p(\mathbf{X}^o_{\backslash t})}{p(\mathbf{X}^o_{\backslash t}|\mathbf{Z}_t)} N \right] + \log(N) \\
&\geq -\mathbb{E}_X \log \left[\frac{p(\mathbf{X}^o_{\backslash t})}{p(\mathbf{X}^o_{\backslash t}|\mathbf{Z}_t)} N \right] \\
&\geq -\mathbb{E}_X \log \left[1 + \frac{p(\mathbf{X}^o_{\backslash t})}{p(\mathbf{X}^o_{\backslash t}|\mathbf{Z}_t)} (N-1) \right] \\
&= -\mathbb{E}_X \log \left[1 + \frac{p(\mathbf{X}^o_{\backslash t})}{p(\mathbf{X}^o_{\backslash t}|\mathbf{Z}_t)} (N-1) \mathbb{E}_{\mathbf{X}^{o,j}_{\backslash t}} \frac{p(\mathbf{X}^{o,j}_{\backslash t}|\mathbf{Z}_t)}{\mathbf{X}^{o,j}_{\backslash t}} \right] \\
&\approx -\mathbb{E}_X \log \left[1 + \frac{p(\mathbf{X}^o_{\backslash t})}{p(\mathbf{X}^o_{\backslash t}|\mathbf{Z}_t)} \sum_{\mathbf{X}^-_{\backslash t} \in \mathcal{X}^-} \frac{p(\mathbf{X}^-_{\backslash t}|\mathbf{Z}_t)}{p(\mathbf{X}^-_{\backslash t})} \right] \\
&= \mathbb{E}_X \log \left[\frac{\frac{p(\mathbf{X}^o_{\backslash t}|\mathbf{Z}_t)}{p(\mathbf{X}^o_{\backslash t})}}{\frac{p(\mathbf{X}^o_{\backslash t}|\mathbf{Z}_t)}{p(\mathbf{X}^o_{\backslash t})} + \sum_{\mathbf{X}^-_{\backslash t} \in \mathcal{X}^-} \frac{p(\mathbf{X}^-_{\backslash t}|\mathbf{Z}_t)}{p(\mathbf{X}^-_{\backslash t})}} \right] \\
&\approx \mathbb{E}_X \log \left[\frac{\frac{p(\mathbf{X}^o_{\backslash t}|\mathbf{Z}_t)}{p(\mathbf{X}^o_{\backslash t})}}{\sum_{\mathbf{X}^-_{1:T} \in \mathcal{X}^-} \frac{p(\mathbf{X}^-_{1:T}|\mathbf{Z}_t)}{p(\mathbf{X}^-_{1:T})}} \right] \\
&= \mathbb{E}_X \log \left[\frac{f(\mathbf{Z}_t, \mathbf{X}^o_{\backslash t})}{\sum_{\mathbf{X}^-_{1:T} \in \mathcal{X}^-} f(\mathbf{Z}_t, \mathbf{X}^-_{1:T})} \right]
\end{aligned}
\tag{10}
$$

There remain two design choices: i) selection of $\mathcal{X}^-$ and ii) formulation of function f. For i), we use a mini-batch approach that $\mathbf{X}^-_{1:T}$ are chosen from other time series inputs in the same mini-batch. For ii), we adopt the average of cosine similarities along the time axis. Specifically, we define the function f using an alternative representations $\tilde{\mathbf{Z}}_{t'}$ which is obtained by inputting the masked input $q_\phi(\tilde{\mathbf{Z}}_{t'}|\mathbf{X}^o_{\backslash t})$ where $\tau \in \Re$ is a temperature hyperparameter:

$$
f(\mathbf{Z}_t, \mathbf{X}^o_{\backslash t}) = \sum_{t' \in \{1:T\} \backslash t} f(\mathbf{Z}_t, \tilde{\mathbf{Z}}_{t'}) = \sum_{t' \in \{1:T\} \backslash t} \exp\left({\mathbf{Z}_t}^T \tilde{\mathbf{Z}}_{t'} / \tau \right) \tag{11}
$$

Alternatively, we can also define the function for negative samples in a similar way.

$$
f(\mathbf{Z}_t, \mathbf{X}^-_{1:T}) = \sum_{t' \in \{1:T\}} f(\mathbf{Z}_t, \mathbf{Z}^-_{t'}) = \sum_{t' \in \{1:T\}} \exp\left({\mathbf{Z}_t}^T \mathbf{Z}^-_{t'} / \tau \right) \tag{12}
$$

Then $I(\mathbf{Z}_t; \mathbf{X}^o_{\backslash t})$ is lower bounded by

$$
I(\mathbf{Z}_t; \mathbf{X}^o_{\backslash t}) \geq \log \left[\frac{\sum\limits_{t' \in \{1:T\} \backslash t} \exp\left({\mathbf{Z}_t}^T \tilde{\mathbf{Z}}_{t'} / \tau \right)}{\sum\limits_{\mathbf{X}^-_{1:T} \in \mathcal{X}^-_{1:T}} \sum\limits_{t' \in \{1:T\}} \exp\left({\mathbf{Z}_t}^T \mathbf{Z}^-_{t'} / \tau \right)} \right] \tag{13}
$$

We can observe that $I(\mathbf{Z}_t; \mathbf{X}^o_{\backslash t})$ is lower bounded by the similar form of NT-Xent objective function in [15] and specifically *in*-version of the supervised contrastive loss in [7].